AFRL-HE-WP-TR-2000-0116

UNITED STATES AIR FORCE RESEARCH LABORATORY

PILOT SELECTION METHODS

Thomas R. Carretta

HUMAN EFFECTIVENESS DIRECTORATE
CREW SYSTEM INTERFACE DIVISION
WRIGHT-PATTERSON AFB OH 45433-7022

Malcolm James Ree

CENTER FOR LEADERSHIP STUDIES
OUR LADY OF THE LAKE UNIVERSITY
411 SW 24TH STREET
SAN ANTONIO TX 78207-4689

SEPTEMBER 2000

INTERIM REPORT FOR THE PERIOD JANUARY 1999 TO JULY 2000

Human Effectiveness Directorate
Crew System Interface Division
2255 H Street
Wright-Patterson AFB OH 45433-7022

NOTICES

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Please do not request copies of this report from the Air Force Research Laboratory. Additional copies may be purchased from:

National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22161

Federal Government agencies and their contractors registered with the Defense Technical Information Center should direct requests for copies of this report to:

Defense Technical Information Center
8725 John J. Kingman Road, Suite 0944
Ft. Belvoir, Virginia 22060-6218

DISCLAIMER

This Technical Report is published as received and has not been edited by the Air Force Research Laboratory, Human Effectiveness Directorate.

TECHNICAL REVIEW AND APPROVAL

AFRL-HE-WP-TR-2000-0116

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

MARIS M. VIKMANIS
Chief, Crew System Interface Division
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 2000	Interim Report - January 1999 to July 2000

4. TITLE AND SUBTITLE

Pilot Selection Methods

5. FUNDING NUMBERS

PE - 62202F
PR - 1123
TA - B1
WU - 01

6. AUTHOR(S)

Thomas R. Carretta*
Malcolm James Ree**

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory*
Human Effectiveness Directorate
Crew System Interface Division
Air Force Materiel Command
Wright-Patterson AFB, OH 45433-7022

Center for Leadership Studies**
Our Lady of the Lake University
411 SW 24th Street
San Antonio, TX 78207-4689

8. PERFORMING ORGANIZATION

AFRL-HE-WP-TR-2000-0116

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory*
Human Effectiveness Directorate
Crew System Interface Division
Air Force Materiel Command
Wright-Patterson AFB, OH 45433-7022

10. SPONSORING/MONITORING

11. SUPPLEMENTARY NOTES

Air Force Research Laboratory Technical Monitor: Dr. Thomas R. Carretta, (937) 656-7014; DSN 986-7014

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This report consists of eight parts. The first part is concerned with describing pilot selection, why it is important, and the knowledge, skills, abilities, and other characteristics typically considered during selection. Part two introduces the concept of validity and the steps involved in doing a validation study. Part three reviews some common methodological issues that make the interpretation of pilot selection studies more difficult and offers "best practices" advice for researchers and practitioners. Part four describes several common criterion measures of pilot training and job performance and research regarding the development of models of performance. Parts five and six review military and commercial pilot selection practices. Where available, information about the construct and predictive validity of the selection methods is provided. Part seven examines future trends in the measurement of pilot aptitude. Finally, part eight provides recommendations for pilot selection researchers and practitioners.

Most important in conducting pilot selection research is scientific rigor. Without scientific rigor, results may be worse than meaningless leading to counterproductive practice. Before setting out to develop a pilot selection system, it is imperative to have a firm foundation in the published literature of human abilities, reliability, validity, job performance measurement, and meta-analysis. Cumulative research results should guide practice.

The military has a long history of research in the selection of pilots and other aviation occupations. In general, they have used both paper-and-pencil tests and apparatus tests such as psychomotor. Cumulative results suggest that general cognitive ability (g) has been a mainstay of military testing and will likely remain so. Measures of pilot job knowledge and psychomotor ability have demonstrated incremental validity when used with measures of g.

14. SUBJECT TERMS			15. NUMBER OF PAGES
Military Pilot Selection	Commercial Pilot Selection	Methodological Issues	71
Meta-Analysis	Group Differences	Structure of Ability	16. PRICE CODE
Causal Models	Validity and Validation Studies		

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev 2-89) Prescribed by ANSI Std Z-39-18
298-102 COMPUTER GENERATED

This page intentionally left blank.

TABLE OF CONTENTS

FIGURES

TABLES

EXECUTIVE SUMMARY

This report consists of eight parts. The first part is concerned with describing pilot selection, why it is important, and the knowledge, skills, abilities, and other characteristics typically considered during selection. Part two introduces the concept of validity and the steps involved in doing a validation study. Part three reviews some common methodological issues that make the interpretation of pilot selection studies more difficult and offers "best practices" advice for researchers and practitioners. Part four describes several common criterion measures of pilot training and job performance and research regarding the development of models of performance. Parts five and six review military and commercial pilot selection practices. Where available, information about the construct and predictive validity of the selection methods is provided. Part seven examines future trends in the measurement of pilot aptitude. Finally, part eight provides recommendations for pilot selection researchers and practitioners.

Most important in conducting pilot selection research is scientific rigor. Without scientific rigor, results may be worse than meaningless leading to counterproductive practice. Before setting out to develop a pilot selection system, it is imperative to have a firm foundation in the published literature of human abilities, reliability, validity, job performance measurement, and meta-analysis. Cumulative research results should guide practice.

The military has a long history of research in the selection of pilots and other aviation occupations. In general, they have used both paper-and-pencil tests and apparatus tests such as psychomotor. Cumulative results suggest that general cognitive ability (*g*) has been a mainstay of military testing and will likely remain so. Measures of pilot job knowledge and psychomotor ability have demonstrated incremental validity when used with measures of *g*.

American law requires job analyses for the development of job selection tests. The results of the analyses should be converted to good practice guided by cumulative knowledge. There is no single ideal pilot selection system, because not all pilots are hired the same way.

In commercial aviation, some pilots are hired directly from the military with many flying hours, some from other airlines, and some directly from training. Although different, all the selection systems should be expected to have three common measurement elements: cognitive ability, conscientiousness (or possibly "integrity"), and job knowledge (Schmidt & Hunter, 1998).

There is a dearth of studies reported by American commercial airlines. Most likely, this is a consequence of two factors: legal liability and competitive edge. Results from two recent surveys by the FAA suggest that US commercial airlines rely heavily on recruiting applicants with prior pilot experience. Prior experience can be assessed in many ways including background checks, interviews, examination of logbooks, flight simulators, and check flights. Aptitude and

personality testing have received relatively little emphasis. In the instances where airlines employ *ab initio* selection (e.g., Bartram & Baxter, 1996), test batteries similar to those commonly found in military pilot selection are used.

The role of psychological evaluation in the licensing of airline pilots has been raised and debated in Europe. Proponents of psychological assessment for licensing see it as a means of identifying psychological deficits of pilots and reducing potential risks to aviation safety. Opponents express fears of abuse and concerns with the use of tests in circumstances for which they were not designed. Clearly, this is an area that will receive attention from aviation industry representatives, aviation psychologists, and pilots for some time.

There has been little use of personality assessment in the United States and the United Kingdom. Personality assessment is more prevalent in continental Europe and the cumulative research suggests that incremental validity could be achieved by using measures of personality, particularly conscientiousness (Barrick & Mount, 1991) or integrity (Schmidt & Hunter, 1998).

PREFACE

This effort was performed under work unit 1123-B1-01 in support of USAF aircrew selection and classification. An abridged version of this report is expected to be published in 2001 as Carretta, T. R., & Ree, M. J. (in press). Pilot selection methods. In B. H. Kantowitz (Series Ed.) & P. S. Tsang & M. A. Vidulich (Vol. Eds.). *Human factors in transportation: Vol. xx. Principles and practices of aviation psychology.* Mahwah, NJ: Erlbaum. .Send e-mail for Dr. Thomas R. Carretta to thomas.carretta@wpafb.af.mil. Send e-mail for Dr. Malcolm James Ree to mree@stic.net.

PILOT SELECTION METHODS

"The quality of the box matters little. Success depends upon the man who sits in it."
-- Baron Manfred von Richthofen, The Red Baron.

INTRODUCTION

This report consists of eight parts. The first part is concerned with describing pilot selection, why it is important, and the knowledge, skills, abilities, and other characteristics typically considered during selection. Part two introduces the concept of validity and the steps involved in doing a validation study. Part three reviews some common methodological issues that make the interpretation of pilot selection studies more difficult and offers "best practices" advice for researchers and practitioners. Part four describes several common criterion measures of pilot training and job performance and research regarding the development of models of performance. Parts five and six review military and commercial pilot selection practices. Where available, information about the construct and predictive validity of the selection methods is provided. Part seven examines future trends in the measurement of pilot aptitude. Finally, part eight provides recommendations for pilot selection researchers and practitioners.

What is Pilot Selection and Why is it Important?

Organizations need people to serve in various capacities and people need jobs. In military aviation, the major goal is achieving and maintaining a high level of mission-readiness. To do so, enough qualified pilots must be available to accomplish mission requirements. This is done by training new pilots, improving retention of experienced pilots, and providing sufficient on-the-job training to achieve mission readiness. Other organizational goals in military aviation include reducing training costs (e.g., lower training attrition, reduce training requirements), avoiding loss of aircraft/loss of life, and achieving diversity in the work force. In commercial aviation, organizational goals emphasize public safety, low training and operating costs, and customer satisfaction. Cascio (1982) provides several examples of the impact of personnel selection on training costs and organizational productivity. In the US Air Force, estimates of the cost of each person who fails to complete undergraduate pilot training range from $50,000 (Hunter, 1989) to $80,000 (Siem, Carretta, & Mercatante, 1988). Obviously, even a small reduction in training attrition could result in large cost avoidance savings.

To achieve these goals, the needs of the organization and of the job applicants must be matched. The process of personnel selection approaches the matching problem from the organization's perspective while taking into account the applicants' view. Personnel specialists identify the employer's needs and select job applicants whose abilities, interests, and characteristics best fit the employer's needs (Guion, 1976). Prior to making a hiring decision, employers try to get a good understanding of each applicant's potential to contribute to achieving the organization's goals. The key to achieving organizational success is the early identification of candidates with the best chance of reaching the required standards for final qualification. Making

the right selection decisions can reduce training costs, improve job performance, and enhance organizational effectiveness.

What Characteristics are Needed to be a Good Pilot?

Since World War I, personnel specialists in both military and commercial aviation have spent a great deal of time, money, and effort attempting to identify the characteristics needed to be a good pilot and the means to accurately measure those characteristics. The military has gone even further, attempting to determine whether a pilot would be better suited to fly fighter or non-fighter aircraft (Carretta, 1989).

In military aviation, pilot applicants typically have little or no prior flying experience. Further, they may not have had prior exposure to the military. Commonly used selection factors include measures of ability (e.g., standardized test scores, college grade point average and major), medical qualification, indicators of "officership" (e.g., commander's ratings from an officer training program), and prior flying experience (e.g., number of hours flown, private pilot's license). Personality assessment is done in some military organizations (e.g., psychological interview), but is less common.

Some commercial airlines have *ab initio* (from the beginning) training programs, where carefully selected applicants with little or no flying experience are put through intensive pilot training courses. However, most commercial airlines prefer to hire experienced pilots to avoid the time and expense of training. When selecting applicants for *ab initio* training, indicators of ability (i.e., trainability) are emphasized. When selecting from experienced pilots, commercial carriers tend to emphasize indicators of prior experience (e.g., certificates and licenses, log book hours, military pilot experience, recommendations) and flying competence (e.g., check flight performance, simulator performance). Military and commercial pilot selection practices will be discussed in greater detail later in the chapter.

Example of Pilot Selection Process

Selection into a military or civilian pilot training program typically is a multi-stage process. Multi-stage selection is the process in which decisions are made at several points. The first stage might be an evaluation of credentials such as flight hours and letters of recommendation. A second stage might be a written test, an interview, or a simulator flight evaluation. The third stage might be flying an aircraft and the last stage might be a final interview. This is different than a selection process in which all the applicant data are collected simultaneously and a single decision made.

Weeks and Zelenski (1998) identified nine barriers to entry into US Air Force pilot training. Barriers included demonstration of minimum educational achievement, interest in the military, interest in the Air Force, officer qualification, officer selection, desire to fly, flying training qualification, pilot training selection, and successful completion of flight screening. Weeks and Zelenski noted that the order of overcoming these barriers varies across individuals. For instance, some individuals may know at an early age that they wish to become a pilot. This occupational choice then drives subsequent choices regarding education and military service.

Others choose a career with the military as a means to finance their education, and only afterward decide to become an officer and pursue a career as a pilot. Another example is the timing and role of flight screening programs. In some instances, pilot applicants attend flight screening after being chosen to enter pilot training. In others, flight screening occurs at an earlier stage, perhaps prior to completion of an officer-commissioning program. Although qualification standards and selection methods vary widely, these barriers are representative of many military pilot selection programs.

An analogous set of barriers could be proposed for commercial pilot selection. For commercial aviation programs involving training, the sequence would be similar to that described by Weeks and Zelenski (1998) with the exception of the military-specific barriers (e.g., officer qualification). Many commercial pilots are former military pilots (Hansen & Oster, 1997), so the Weeks and Zelenski framework could be a starting point. Additional stages for former military pilots applying for commercial pilot jobs could be added regarding interest in commercial aviation, pilot qualification, and pilot selection.

Are Effective Pilots "Selected" or "Trained?"

Both selection and training play important roles in producing effective pilots who will allow the organization to meet its goals. Effective selection procedures will produce cost-avoidance savings through reduced attrition and reduced training requirements and will lead to improved job performance and improved organizational effectiveness. Poor selection will result in increased training attrition, training requirements, and cost (i.e., more flying hours needed to train poor applicants to achieve some standard), and will lead to poor job performance and poor organizational effectiveness. Effective training methods can help reduce training attrition and contribute to improving organizational effectiveness (Patrick, in press; Smallwood & Fraser, 1995; Walter, 1998).

VALIDITY AND VALIDATION STUDIES

What is Validity?

Validity is the most fundamental testing and selection issue. As described by Jensen (1980) "A test's validity is the extent to which scientifically valuable or practically useful inferences can be drawn from the scores." (p. 297). However much effort is made to develop selection methods based on theories of the relations between personnel characteristics and performance, they will come to nothing without validity. Theory without proof is, at best, worthless. Frequently, theories without proof divert resources and hinder advancement.

Historically, we have acknowledged three types of validity: content, construct, and criterion (predictive). A test has *content validity* to the extent that its items are judged to represent some clearly specified area of knowledge, skill, ability, or characteristic. This judgment is often based on the consensus of subject-matter-experts (SMEs). For example, psychologists might be SMEs for making judgments about tests of cognitive processes (e.g., intelligence,

memory, processing speed) whereas experienced pilots might be appropriate SMEs for measures of flying job knowledge or performance.

Whereas content validity is based on expert judgment, *construct validity* is concerned with the scientific attempt to determine what a test actually measures. Construct validity becomes an important issue when we have some theory about the nature of the trait that we measured. A theoretical foundation allows us to develop and test hypotheses about what will happen under specified conditions. A test is said to have construct validity if it predicts behavior in specific situations that would be inferred from our theory.

Criterion or predictive validity is the ability of test scores to predict performance in some activity (criterion) external to the test itself. Typically, in personnel selection the criterion consists of one or more measures of training or job performance. Though content and construct validity are very desirable for enhancing our understanding of the tests and criteria, neither is essential for criterion validity. All that is needed for criterion validity is that the test predicts the criteria. An important concept related to criterion validity is incremental validity. A test has incremental validity if it improves prediction of the criteria beyond that provided by some baseline test. Although all three types of validity are important in personnel measurement and selection, criterion validity will be emphasized here, due to its greater use in pilot selection.

What is a Validation Study?

The description is based on best professional practice and legal requirements. The legal requirements come from case law especially Griggs v. Duke 1971 and from the federal *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, 1978). The general standards for validity studies are described in § 1607.5 of the *Uniform Guidelines on Employee Selection Procedures.*

Selection necessarily implies screening of job applicants and rejection of some. As noted by Jensen (1980), there are two justifications for selection. The first is when the pool of applicants is larger than the number of training or job positions. The second is when the predictive validity of the selection procedures can be demonstrated. Tests or other selection methods (e.g., biodata, interviews, recommendations) are said to have predictive validity to the extent that they would distinguish between the performance of selectees and rejectees if all of them had been selected.

In their pursuit of an ideal pilot selection system, aviation psychologists have examined a variety of personnel constructs and measurement methods (Hunter & Burke, 1995). A formal validation study is required to determine the utility of these constructs and methods for predicting training and job performance. Guion (1976) describes a four-step procedure for forming and testing hypotheses about personnel selection.

Perform a job analysis. The first step is to identify important job performance constructs, usually through job analysis (Cascio, 1991; Gael, 1988; McCormick, 1976, 1979). The goal of job analysis is the establishment of job, task, and cognitive requirements or Knowledge, Skills, Abilities, and Other (KSAO) requirements. It can be accomplished many different ways. Cascio

(1991) provides a good discussion of the methods. Results from the job analysis can lead to the development of a structural taxonomy and specification of predictor and criteria measures.

Develop operational definitions of important constructs. The second step is to develop operational definitions of these job performance constructs and ensure that they show acceptable construct validity. As previously discussed, construct validity is based on theory and is determined by testing hypotheses about the relations between the tests and performance criteria. Construct validity of a psychomotor test could be examined by administering it along with marker tests whose properties are well known and examining the relations between the psychomotor test and marker tests.

Identify a set of predictors and criteria. The third step is to propose a set of predictor and criteria variables. The choice of predictors should be guided by theory. They should be developed using psychometric techniques to insure appropriate content, difficulty, and precision of measurement. Pilot job performance criteria must be established using the same psychometric guidelines of proper content, difficulty, and precision of measurement. The criterion is usually some measure of occupational performance such as training completion or accomplishments, hands-on job ratings, work samples, job knowledge, supervisor ratings or measures of productivity. Examples include supervisory ratings, accident reports, or direct indicators of job performance such as percent of on-time arrivals or percent of enemy targets destroyed. Performance ratings are the most frequently used criterion measure (Pulakos, 1997). In practice, most criterion variables are positively correlated. This means that most criterion variables measure aspects of the same underlying construct and the main feature of criterion development is specifying a sufficient number of measures, avoiding criterion contamination from extraneous features, and covering the breadth of the criterion construct.

Examine predictive validity. In the fourth and final step, select predictor and criterion measures and examine predictive validity. When the predictor and criterion measures have been deemed suitable, they can be administered to an appropriate sample in either a predictive or a concurrent validation design. In a predictive design, the appropriate sample is a large group of applicants. The predictor measures are administered during application and the criteria are collected after those selected have completed training or been on the job for some period such as three, six, or twelve months. In a concurrent design, a large sample of job incumbents is administered the predictor and the criterion measures simultaneously (concurrently). In both validity designs, the criteria data are available for only a sample of those selected for training or employment. This leads to a selected sample and the artifact of range restriction as described later in this chapter.

During validation, the data are analyzed and inferences are drawn from the scores on the predictor and criterion measures. The index used to assess predictive validity is usually the correlation coefficient (r) or the multiple correlation (R) if there is more than one predictor. The Griggs v. Duke 1971 decision established the commonly used $p < .05$ significance level (Type I error rate). The final part of the analysis is the reporting of the validity study and its results. Whetzel and Oppler (1997) provide a good introductory overview of this process.

Other considerations. In addition to the selection system's utility for identifying those likely to be successful, there are other important considerations in personnel selection. These include whether or not the selection methods predict training and job performance equally well for members of different sex and ethnic/racial groups (i.e. predictive bias) and whether or not a test differentially disqualifies members of different subgroups (i.e., adverse impact).

COMMON METHODOLOGICAL ISSUES IN PILOT SELECTION

Courses in research methods and statistics are common for personnel specialists and there are established guidelines for conducting studies of personnel measurement and selection (APA, AERA, & NCME, 1985; SIOP, 1987). Despite this, many studies of personnel measurement and selection embody methodological issues that cloud their interpretability. We have identified five major methodological issues that can influence conclusions about construct and criterion-related (predictive) validity. There are more, but these five are common and lead to incorrect conclusions and decisions about the effectiveness of pilot selection systems. The five issues are misunderstanding constructs, misinterpretation of factor analytic results, lack of statistical power, failure to estimate cross-validation effects, and misinterpretation of correlations and regression. Our purposes in discussing them in this chapter is to raise the readers' awareness so they will be able to read the published literature with a critical eye and to offer "best practices" solutions for researchers and practitioners. Carretta and Ree (in press) provide a more detailed discussion.

Misunderstanding Constructs

Abstractions such as "airmanship," "intelligence," "situational awareness," or "workload" are called constructs. They cannot be observed directly and must be inferred from some test, measurement scale, or questionnaire that operationalizes the components of the construct. There is no scientific value to a construct that cannot be measured.

It is important to remember that a construct can be measured by many means. Hunter and Hunter (1984) have demonstrated this and Spearman (1927) noted it in his idea of "the indifference of the indicator" as Jensen (1993) has pointed out. Walters, Miller, and Ree (1993) have labeled the specious reasoning that because two tests look different they must measure different constructs, the "topographical fallacy." The appearance of the items or tasks of a test is not a reliable indicator of what is being measured. Results of a construct validation study provide a good indicator of what is being measured.

Consider the following example. Several NATO countries use interviews as part of their military pilot selection process (Hansen, 1999). On the basis of the topographical fallacy (i.e., differing appearances), the U. S. Air Force (USAF) considered using a structured interview in their pilot selection process (Walters et al., 1993). Structured interviews have predetermined rules for obtaining, observing, and evaluating responses. In a recent review of the literature on employment interviews, Whetzel and McDaniel (1997) concluded that structure tends to make the interview more reliable and valid. In their meta-analytic review of the utility of personnel selection methods, Schmidt and Hunter (1998) concluded that structured interviews were incrementally valid to measures of general cognitive ability for predicting training and job

performance. Walters et al. (1993) reported that the highly structured USAF pilot selection interview measured educational background, motivation to fly, self-confidence and leadership, and flying job knowledge. Additionally, three broad ratings of probable success in pilot training, bomber-fighter flying, and tanker-transport flying were made. The sample was 223 USAF pilot trainees who were administered the structured interview, Air Force Officer Qualifying Test (AFOQT; Skinner & Ree, 1987), and computer based cognitive "information processing" (verbal classification, mental rotation, and short-term memory) and personality (self-confidence and attitudes toward risk) tests. Passing-failing pilot training was the criterion. The seven structured interview scores had an average validity of .21. When the seven scores were added to regression equations containing the AFOQT scores (subtest average validity of .28) and the computer-based test scores (average validity of .18), no incremental validity was found. Despite the difference in appearance between the paper-and-pencil AFOQT and the interview, the interview was not able to account for any unique prediction of pilot performance.

It is suggested that researchers and practitioners in pilot selection administer their selection instruments along with a battery of tests of known constructs to have a better understanding of what is being measured by their selection instruments.

Misinterpretation of Factor Analytic Results

Factor analysis (Spearman, 1904, 1927) is the name given to a group of statistical techniques for determining the latent or unobservable sources of variance in a correlation matrix. In personnel selection, it is commonly used to identify the constructs measured by a test battery or by job performance criteria. Rotation of the initial factor solution plays an integral role in factor analysis and can produce a problem in interpretation. During rotation the variance from the first factor is spread across all the factors. This can make the first factor seem to disappear when in reality it has simply been distributed across all the factors. The solution is to either not rotate or use a residualized (Schmid & Leiman, 1957) hierarchical solution.

Lack of Statistical Power

Statistical power is the probability of detecting a significant effect, such as a difference between means or a non-zero correlation, when present. Specifically, it is the probability of rejecting a false null hypothesis (Cohen, 1987). Although the topic is covered in most introductory statistics classes, relatively few published studies report power for the test statistics (such as F, t, or z,) used. See for example Ree and Earles (1991, 1993) and Walters et al. (1993). Sedlmeier and Gigerenzer (1989) reported two surveys of a prestigious applied psychology journal and showed that the average statistical power for published studies was only .46 and fell to .37 two decades later. This means that the researchers could only detect an existing effect 46 percent (or 37 percent) of the time. On the other hand, 54 percent of the time (or 63 percent of the time) they would fail to find the existing effect! No researcher should conduct a study with less than a high chance of detecting a significant result. Low statistical power makes it more likely that we will draw incorrect conclusions from our studies. Too many pilot selection studies have been performed with small samples that inevitably yield low statistical power.

Cohen (1987), noted that statistical power is a joint function of the degree to which the sample values reflect their true values in the population (i.e., the reliability of the sample values), effect size, Type I error rate (significance level), and sample size. Before conducting a study, power tables (Cohen, 1987) should be consulted. An informative discussion of sample size requirements (see especially page 222) is given by Schmidt and Hunter (1978). Schmidt, Hunter, and Urry (1976) provide tables showing sample sizes necessary for sufficient statistical power in validation studies given varying selection ratios, reliabilities, and effect sizes .A single rule-of-thumb cannot be given. In general, the higher the selection ratio and the less reliable the variables are, the larger the sample must be.

Failure to Cross-Validate

Cross-validation refers to the use of a regression equation computed in one sample being applied in another sample. This is done to determine the extent to which the initial regression solution can be expected to generalize to another sample. The stability and generalizability of the regression solution are important because the results of the regression equation will be used to make personnel selection decisions about people who were not in the original validation sample. In general, the correlation of the several predictors with the criterion will go down in this second or cross-validation sample (Wherry, 1975). This reduction of the correlation is called shrinkage from overfitting.

The classic paradigm for cross-validation was provided by Mosier (1951) in which a single sample is drawn from a population and then divided into separate validation and cross-validation samples. Murphy (1983) has pointed out that the correlation in the cross-validation sample is still a consequence of overfitting, as there was only one sampling from the population. Moreover, even if there were two samplings from the population, the validation and cross-validation multiple correlations would be only two values out of a virtually infinitely large set of values. That is why we recommend the use of Stein's operator (Stein, 1960). Kennedy (1988) demonstrated the accuracy of Stein's formula for estimating the mean of the distribution of all possible cross-validated correlations from the population from which the sample was selected. Stein's operator has the advantage of allowing estimates on the largest available sample while offering an estimate of cross-validity.

Misinterpretation of Correlations and Regression

Selection studies use correlation and regression as a general model and analytic technique. Predictors of success in pilot job performance are correlated with measures of success in pilot job performance. There is an extensive literature on correlation and regression (Cronbach, 1971; Messick, 1989). In the next sections we discuss several issues that can lead to the misinterpretation of correlations and regression, and provide solutions.

Holding job experience constant. Ability research is generally correlational and the interpretation of correlations can be fraught with hazards. As an example, consider the correlation of an ability test and ratings of pilot job performance. "Artificially" low correlations that could lead to inappropriately abandoning the ability test can occur for a variety of reasons including the effects of range restriction, unreliability, and the influence of moderating variables.

Range restriction and unreliability will be discussed later in this section. It is noted in the *Principles for the Validation and Use of Personnel Selection Procedures* (SIOP, 1987) that the relationship between ability (or any other measure) and occupational criteria is best represented with the effect of job experience removed. This can be done easily by using partial correlation and "partialing-out" experience from the relationship between ability and the occupational criteria. Carretta, Perry, and Ree (1996) provided an example. When they correlated ability test scores with ratings of situational awareness (SA) for 171 F-15 pilots, the observed correlation of ability and SA was .10. However, when F-15 flying experience was partialed-out, the correlation was .17, an increase of 70% in predictive efficiency. It would have been incorrect to report the correlation of ability and SA as .10.

The idea of partial correlation can be subsumed under "mediation," which means that one variable acts through another variable to exert its influence on a third variable. For instance, "A →B→ C" indicates that variable A acts through variable B to exert its influence on variable C. In this instance there is no direct influence of A on C and we do not specify "A→C." This does not mean that variable A has no influence on variable C, but rather that A works through B to influence C. Hunter (1986) provided an informative model of mediation in the area of job performance. He demonstrated that job knowledge mediated the relationship between ability and job performance for numerous jobs. Ree, Carretta, and Teachout (1995) illustrated this mediation for pilot trainees (see Figure 1). Ree et al. examined the influence of general cognitive ability (*g*) and prior job knowledge (JK_P) on the acquisition of job knowledge acquired during early, middle, and late pilot training (JK_{T1} to JK_{T3}) and early (T-37) and late (T-38) hands-on flying performance (WS_1 and WS_2). In their study, general cognitive ability (*g*) had both direct and indirect influences on the acquisition of aviation job knowledge and hands-on flying performance during pilot training. It is necessary to partial-out the effect of job experience to know the true relationship of a predictor to job performance.

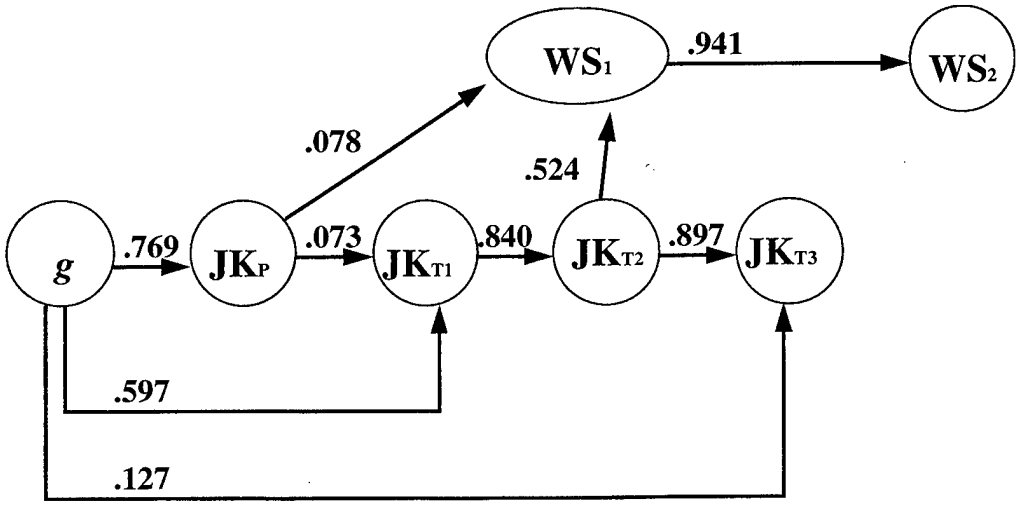


Figure 1. Ree, Carretta, and Teachout (1995) model of the influence of general cognitive ability (*g*) and prior job knowledge (JK_P) on the acquisition of additional job knowledge (JK_{T1} to JK_{T3}) and sequential training performance (WS_1 and WS_2).

Range restriction. Studies of training and job performance frequently use censored samples. When the variance of one or more variables has been reduced due to prior selection, censoring occurs. Range restriction is the name given to this reduction in variance. For example, military organizations do not admit all those who apply and commercial airlines typically do not hire all pilot applicants. In military pilot selection, applicants may have been screened on the basis of aptitude test scores, completion of a college degree, completion of an officer-commissioning program, medical fitness, prior flying experience, selection interview, and vocational interest. Censored range-restricted samples have been shown to cause artifacts that may lead to erroneous conclusions (Morrison & Morrison, 1995) and to inappropriately abandoning predictive measures.

Range-restricted samples can produce estimates of correlations that are artificially substantially less than they would be in an uncensored sample (Martinussen, 1997; Thorndike, 1949). In some instances, correlations based on censored samples can even change signs from their population value (Ree, Carretta, Earles, & Albert, 1994; Thorndike, 1949).

Damos (1996) argues against the use of corrections for range restriction, noting that organizations do not administer selection tests to a completely unscreened population. She contends that the uncorrected correlation provides the most accurate estimate of the strength of the relationship between two variables. The following example disagreed.

A dramatic illustration of the detrimental effects of range restriction was provided by Thorndike (1949, pp. 170-171). An experimental group of 1,036 US Army Air Corps aircraft pilot applicants was admitted to training without regard to their scores on five aptitude tests during the Second World War. Correlations were computed with the training criterion for all participants (n = 1,036) and for those pilot candidates (n = 136 out of 1,036) that would have been selected had the strict standards in effect been used. In the range restricted sample (the 136 qualified pilot candidates) the average decrease in the five validity coefficients was .29. The Pilot Stanine composite derived from the five tests had a correlation of .64 with training outcome in the unrestricted sample. In the range-restricted sample it dropped to .18. The most dramatic change from the unrestricted to the range-restricted sample occurred for a psychomotor test where the correlation changed sign from +.40 to -.03. It is clear that the validity estimates were adversely affected by range restriction. Further, wrong decisions would have been made as to which tests to implement had only the range-restricted correlations been reported.

Range restricted samples are commonplace in pilot selection research. Goldberg (1991) observed that "... one can always filter out or at least greatly reduce the importance of a causal variable, no matter how strong that variable, by selecting a group that selects its members on the basis of that variable" (p. 132). He noted that the more restricted the variance of a variable, the less its apparent validity.

Statistical corrections for range restriction are available and should be applied to provide better statistical estimates. "Univariate" corrections described by Thorndike (1949) are appropriate if censoring has occurred on only one variable. However, the multivariate correction (Lawley, 1943; Ree et al., 1994) is more appropriate if censoring has occurred on more than one

variable. These corrections provide better statistical estimates and tend to be conservative (Linn, Harnish, & Dunbar, 1981). The corrections still tend to underestimate the population values. Johnson and Ree (1994) offered free windows-based software to perform either univariate or multivariate corrections. When working with range-restricted samples, corrections always should be used.

Unreliability of measures. Reliability refers to the accuracy of measurement of a test, interview, or other selection device. Reliability can be estimated by correlation between test forms (test-retest) and under certain conditions from the single administration of a test (internal consistency) (McDonald, 1999). The method that should be used depends on several factors including the content of the test or scale and the question being asked. Internal consistency estimates are appropriate if the items are independent of one another and the question of interest is whether or not the items measure the same construct. Test-retest estimates are appropriate if the test is speeded, the items are not independent of one another (e.g., most psychomotor tests), or the question of interest involves stability of performance over time or across alternate forms.

The use of unreliable measures will lead to incorrect conclusions. (Spearman, 1904). The magnitude of the correlation between variables is limited by their reliabilities. Correcting for the unreliability of variables informs us about the true relationship between predictors and criteria. Correlations between predictors and criteria that change from low to moderate or high after correction suggest that the predictor could help increment validity if it (or the criteria) were more reliable. If validities remain low to moderate after correction for unreliability it is likely that the criterion has other sources of variance that are not being predicted.

Carretta and Ree (1995a) provided an example when they examined the validity of the 16 Air Force Officer Qualifying Test (AFOQT) tests against five USAF undergraduate pilot training criteria. The validities of the AFOQT tests were examined as observed, corrected for range restriction, and corrected for both range restriction and unreliability (of the predictor and criterion). The average magnitude of the correlations of the 16 tests with the five criteria varied from .03 to .13 for the observed correlations, from .10 to .25 for the range-restriction-corrected correlations, and from .25 to .58 for the fully-corrected correlations. The use of appropriate range restriction and unreliability corrections removes artifacts that are inevitable in all studies (Hunter & Schmidt, 1990).

Unreliability also has an effect on regression coefficients. Fuller (1987) provided a mathematical demonstration that unreliability biases b, the estimate of β, the population regression parameter. The estimate is reduced because the less-than-perfect reliability causes the value to be lower than if a more reliable measure had been used.

Unreliability of measures also plays a part in factor analysis. Just as unreliability reduces correlations and regression coefficients, unreliability reduces factor loadings. This causes an underestimate of the true factor loadings. The underestimation can be corrected by dividing the factor loading by the reliability. Ree and Earles (1993) reported results from Jones and Ree (1998) that showed the correlation of factor loadings and test validities to be .78. After correcting the factor loadings for the unreliability of the tests, the correlation was estimated to be .98.

One solution to increase reliability is to add items to unreliable measures (e.g., more test questions, additional job performance ratings, etc.). Other solutions would be to improve instructions and to remove ambiguity from existing items and from scoring. If these remedies are ineffective, it may be necessary to discard the measure. Nonetheless, reliability should be estimated for all predictors and training or job performance criteria.

Dichotomization of criteria. In training studies of pilots or other aviation occupations, it is not unusual to have criteria that have been artificially divided into two categories, pass or fail. Did the student pass or fail the pilot course? Did the mechanic pass or fail the airframe certification test? Dichotomization of the criteria causes correlations to appear lower than they should and places an upper limit on the magnitude of the correlation that depends on the proportion in each of the pass and fail categories. With proportions of 50-50 there is no biasing effect on the correlation, but when the proportions deviate from 50-50, there is a downward bias on the correlation. For example, if a correlation between two variables is .50 before dichotomization and the dichotomized criterion has proportions of 50-50, the correlation in the study will be .50. However, if the proportions are 90-10, 80-20, 70-30, or 60-40, the correlations will be .29, 35, 38, and .39 respectively. If the correlation before dichotomization were .25, the after-dichotomization correlational values for the proportions 90-10, 80-20, 70-30, and 60-40, would be: 15, 17, 19, and .20. This has long been recognized as a problem and a statistical correction for the dichotomization (Cohen, 1983) provides an estimate of the correlation had the variable not been dichotomized.

Examination of effects for subgroups. Sometimes it is possible to confuse between group effects for within group effects. This can lead to a predictor being declared valid when it is the group differences that create the appearance of validity. Consider the following example. It may be necessary to lift a heavy object above your head during aircraft maintenance. In a validation study, a physical lifting test was administered. In validation studies it is common to note that both sexes or two or more ethnic groups were included in the sample. In the combined sex or ethnic groups the correlation between the predictor and criterion may be moderate or large, but within each group the correlation is low or zero. This suggests that the validity in the combined group may be nothing more than a statistical artifact (see for example, Hogan, 1991 and Ree, Carretta, & Earles, 1999). This seeming paradox, high or moderate correlation in the combined group and zero or low correlation in each individual group, is not a psychological phenomenon, but a mathematical consequence of correlation and regression being data-driven. Figure 2 shows three cases where the predictor-criterion relationship is very different in the combined group than in the two subgroups. Figure 2a shows an instance in which there is a slight positive predictor-criterion relationship in each subgroup, but a much stronger positive relationship when the subgroups are combined. In Figure 2b, the correlation in the combined group is zero, but each subgroup the correlation is negative. Finally, in Figure 2c, there is a strong positive correlation in the combined group, a slight negative correlation in one subgroup and a zero correlation in the other subgroup. It is possible to have nearly any combination of three correlations so long as the correlation in the combined group is neither -1 nor +1.

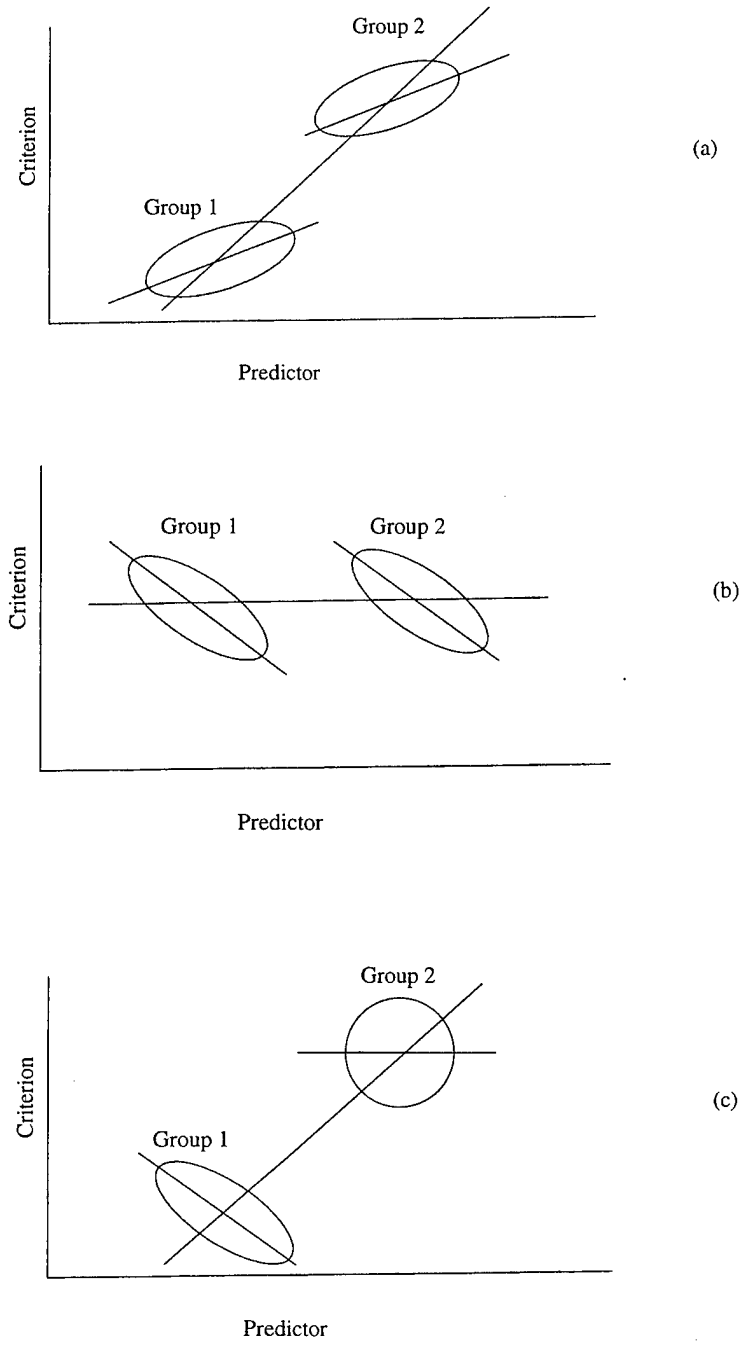


Figure 2. Examples of different predictor-criterion relationships for subgroups and combined group

Note. Figure 2a shows a weak positive predictor-criterion relationship for each subgroup and a stronger positive relationship for the combined group. Figure 2b shows a negative predictor-criterion relationship in each subgroup, but a zero relationship in the combined group. Figure 2c shows an example of a slight negative relationship in one subgroup, a zero relationship in the other subgroup, but a strong positive predictor-criterion relationship in the combined group.

Ree et al. (1999) provided several examples of this two-group phenomenon and proposed and demonstrated a general hierarchical linear models analysis to address the issue. The first step in the Ree et al. hierarchical linear models analysis approach is to test for equivalence of the variance errors of estimate (Gulliksen & Wilks, 1950; Jensen, 1980; Reynolds, 1982). A series of F tests of specified hierarchical linear models is appropriate if the errors of estimate are equal. The first F test in the hierarchical models series compares a linear model with two slopes and two intercepts with a model with only one slope and one intercept. A non-significant F means that there are no between-groups differences and analyses should be conducted at the within-group (combined groups) level. When significant differences are found, additional tests of the differences between slopes and between intercepts would be performed. In addition to conducting these statistical analyses, it is recommended that researchers plot their data for each subgroup and for the combined group.

This linear models approach is less than optimal when comparing more than two groups. With more than two groups, a Within and Between Analysis (WABA; Dansereau, Alutto, & Yammarino, 1984) is applicable.

Weighting of variables. Typically, aviation job applicants are given a series of tests, or multiple interviews, or a simulation task with many scores, or a combination of all of these. Then a decision is made by the selecting agency, frequently combining the scores and other applicant information by addition to form a composite. Frequently, the various parts of the composite will be given greater importance by weighting them more. The score on the composite, rather than its parts, will be used to make a decision. Variable weighting to create composites has been the subject of both analytic and empirical study. Two common weighting methods include unit weighting and criterion-based regression weighting. Unit weighting assigns each score the same weight and simply adds the scores together to create a composite. Criterion-based regression weighting uses the best-fitting weights when some criterion is regressed on a set of predictors. Criterion-based regression weighting is used frequently in pilot selection (Walters et al., 1993), even though several studies argue for unit or simple weighting. Three decades ago, Aiken (1966) thought the controversy over the use of simple weights was settled and was surprised to find colleagues arguing for regression-based weights on an intuitive basis. Two decades ago, Wainer (1976, 1978) showed only small losses in predictive efficiency from equal weights when compared with regression weights. He noted that selection usually involved ranking and top-down selection rather than predictive efficiency, making weighting schemes of little importance. Wilks (1938) proved a mathematical theorem showing that under very common circumstances, almost all weighted composites of a set of variables are very strongly correlated. In other words, if two different sets of weights were applied to a set of variables to produce two composites, the correlation between the two composites will be very high.

Ree, Carretta, and Earles (1998) demonstrated the consequences of Wilks' (1938) theorem through multiple examples. They also provided numerous examples from published studies showing near identical rankings for composites based on various weighting schemes (e.g., unit weights, regression weights, factor weights, and policy-capturing weights).

When considering weighting variables it is sufficient to know which are important and then use simple or unit weights. Considering other than simple or unit weights, Wainer (1976) said it succinctly, "it don't make no nevermind."

Recommendations for Researchers and Practitioners

Ability research is fraught with pitfalls that can lead to incorrect inferences. In this section, we offer recommendations for each of the methodological issues raised earlier. It is worth noting that many or all of them can occur in a single study. The effects will be compounded. Ree (1995) provided an example study with multiple problems as caused by multiple issues. This example study is reminiscent of many others and shows how incorrect conclusions can occur.

We recommend the following to avoid these problems:

1. Use reference tests to establish construct validity. The appearance of a test is an unsure indicator of what is actually being measured.

2. Misinterpretation of factor analysis results is often the direct consequence of rotation. The problem of the disappearing first factor as a result of rotation can be avoided by residualized hierarchical factors, or by using unrotated principal components or unrotated principal factors.

3. Take steps to ensure sufficient statistical power. It is wasteful to do studies when you have a low probability of detecting the effect.

4. Estimate cross-validities using one of several non-sampling methods. These methods have the advantage of allowing estimation on the largest available sample while offering an estimate of cross-validity.

5. When interpreting correlations:
   a. Hold job experience constant.
   b. Evaluate the utility of mediators. Some variables exert their influence on others both directly and indirectly through some mediating variable.
   c. Correct correlations for statistical artifacts such as dichotomization of variables, range restriction, and unreliability of measures. Less biased statistical estimates are preferable.
   d. When applicable, examine effects for subgroups as well as the total group. Relationships observed in the total group may be radically different from those observed in subgroups.
   e. Consider simple or unit-weighting schemes to express the relationships among related variables (e.g., aptitude scores, job performance ratings). In many common instances, simple or unit-weighting schemes are as effective as more complex and sometimes costly procedures (e.g., weights derived through policy capturing exercises or statistical procedures).

MEASURES OF PILOT TRAINING AND JOB PERFORMANCE

Common Measures of Pilot Training and Job Performance

Researchers and practitioners in pilot selection spend most of their effort on identifying crucial pilot abilities and characteristics and ways to measure them (e.g., Carretta, Rodgers, & Hansen, 1993). Relatively little attention has been given to the development of measures of pilot training and job performance. The most common measures of performance are based on simple dichotomous scores (Hunter & Burke, 1995). Examples include passing/failing training (e.g., Burke, Hobson, & Linsky, 1997; Walters, Miller, & Ree, 1993), performance above or below some arbitrary performance cutoff point (e.g., Hörmann & Maschke, 1996; Long & Varney, 1975), and fighter/non-fighter recommendation (Weeks, Zelenski, & Carretta, 1996). Less common are alternate training and job performance criteria such as number of flying hours needed to complete training (Duke & Ree, 1996), academic and check flight grades (Carretta & Ree, 1996; Olea & Ree, 1994; Ree, Carretta, & Teachout, 1995), class rank (Carretta, 1992b), simulator grades (Gress & Willkomm, 1996), and peer and supervisory ratings of job performance (Carretta, Hansen, & Woodhead, 2000; Carretta, Perry, & Ree, 1996).

A common complaint is that dichotomous measures such as passing-failing training place an upper limit on validity coefficients (Damos, 1996). Others contend that although traditional flying training performance measures (e.g., passing-failing training, flying grades) provide adequate information about overall performance, they lack sensitivity and fail to reflect specific deficiencies that could be addressed during pilot candidate selection (Carretta et al., 2000).

Even when alternate performance criteria are developed, there is no guarantee that they will lead to a better understanding of the relation between selection factors and performance. Carretta (1992b) examined the relations between various selection factors and performance in USAF undergraduate pilot training. Training criteria included a dichotomous passing-failing score and four alternative measures of class rank based on a combination of academic grades, daily flying grades, and check flight grades. The class rank scores were shown to be related closely to advanced training recommendation (fighter or non-fighter aircraft). This suggested that the class rank scores were a reasonable measure of pilot candidate quality because fighter assignments are considered more prestigious and demanding than are nonfighter assignments. Despite the differences in the criteria, the passing-failing criterion had an average correlation of .97 with the four class rank criteria. Use of alternate criteria would have had little effect on pilot selection decisions.

Models of Job Performance

It is important to have a model of performance prior to development of training or job performance criteria. Several models have been proposed. Sackett, Zedeck, and Fogli (1988) proposed a model that distinguishes between typical and maximum performance. Sackett (personal correspondence, November 20, 1996) suggested that typical performance is a function of ability and conscientiousness.

Campbell, McHenry, and Wise (1990) proposed a job performance model with eight dimensions. The dimensions are: 1) job-specific task proficiency, 2) non-job-specific task proficiency, 3) written and oral communication task proficiency, 4) demonstrating effort, 5) maintaining personal discipline, 6) facilitating peer and team performance, 7) supervision/leadership, and 8) management/administration. Campbell, McCloy, Oppler, and Sager (1993) described these dimensions as the highest-order factors that could be useful and discounted the existence of a general job performance factor. They contend that "... three of the factors -- core task proficiency, demonstrated effort, and maintenance of personal discipline -- are major performance components of every job." (pp. 48-49, emphasis in the original). They also noted that not all of the factors are relevant to all jobs.

Campbell et al. (1990) evaluated their model on nine US Army entry-level jobs and found five factors. Ree and Carretta (1998) subsequently used Campbell et al.'s correlations of the factors to conduct a confirmatory factor analysis. Two higher-order factors emerged that correlated .39. The first of these higher-order factors was composed of Campbell et al.'s lower-order factors of "core technical proficiency" (job-specific task proficiency) and "general soldering proficiency" (non-job-specific task proficiency). The second factor was composed of Campbell et al.'s other three lower-order factors of "effort and leadership" (demonstrating effort), "personal discipline" (maintaining personal discipline), and "physical fitness and military bearing." The first of these factors was interpreted as a "can do" factor and the second as "will do." The correlation between the two higher-order factors (.39) suggests some third-order factor as a common source.

Similarly, Lance, Teachout, and Donnelly (1992) modeled the latent structure of job performance using hierarchical confirmatory factor analysis. They observe four lower-order factors (job proficiency dimensions) with a mean correlation of .42. An Eigenvalue analysis indicated that a common first factor accounted for 59% of the variance among these lower-order factors. This suggests a hierarchical structure, perhaps with a general factor.

Sometimes, despite the intention to develop multidimensional criteria, unidimensional measures are actually produced. For example, Houck, Whitaker, and Kendall (1991) performed a job analysis that resulted in 31 behavioral items considered to represent eight performance dimensions for military fighter aircraft pilots. Rating items represented general traits, tactical game plan, system operation, communication, information interpretation, tactical employment – beyond visual range (BVR) weapons, tactical employment - visual maneuvering, and tactical employment - general. Even though trained psychologists and pilot subject-matter-experts participated in the writing of criterion items, a single performance factor accounted for 92.5% of the variance in performance ratings for a sample of 171 F-15 pilots (Carretta et al., 1996).

Given these results, it is reasonable to search for a common job performance factor and investigate whether it has a hierarchical relationship to the factors proposed by Campbell et al. (1990) or found by Lance et al. (1992). In a broad-based meta-analysis of the factors of the job performance criteria space, Viswesvaran, Schmidt, and Ones (1996) cumulated results across 297 studies and across many measurement factors and sources. Their results showed a higher-order general factor for job performance. Viswesvaran et al. suggested that the general job

performance factor was both theoretically and practically important, but required additional research.

MILITARY PILOT SELECTION

Historical Overview

Dockeray and Isaacs (1921) reported that Italy, prior to World War I, was the first country with a pilot selection research program. The Italians used measures of reaction time, emotional reaction, equilibrium, perception of muscular effort, and attention. At the same time, the French were investigating reaction time and emotional stability.

In the World War I era, Yerkes (1919) showed that measures of intelligence were valid predictors of pilot training success. Between the wars, Flanagan (1942) noted that the American aviation selection exam was a general mental battery testing comprehension and reasoning.

Most of the World War I research reflected Spearman's (1904) two-factor theory that demonstrated the existence of a general cognitive factor and a test-unique factor. The work of Thurstone (1938) changed the emphasis from Spearman's two-factor theory to the theory of multiple aptitudes (i.e., innate or acquired capabilities; talents). Thurstone's theory was eventually reified in multiple aptitude batteries such as the *Differential Aptitude Tests* (Bennett, Seashore, & Wesman, 1982), *General Aptitude Test Battery* (Dvorak, 1947; Hunter, 1980), *Armed Services Vocational Aptitude Battery* (ASVAB; Earles & Ree, 1992), and the *Air Force Officer Qualifying Test* (AFOQT; Carretta & Ree, 1995a). The AFOQT has played an important role in the selection of US Air Force pilots for more than 40 years.

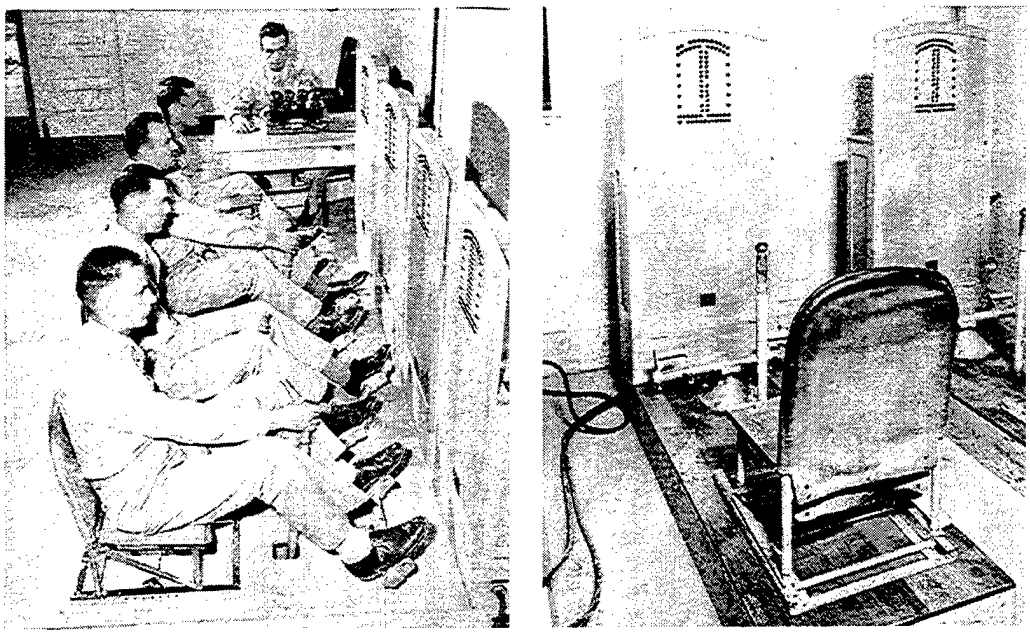


Figure 3. Air Force School of Aerospace Medicine Complex Coordination Test

World War II brought a renewed interest in pilot selection. Influenced by Thurstone's multiple aptitude theory, the American Army (Melton, 1947) and Navy (Fiske, 1947; Viteles, 1945) used several ability measures for pilot selection. These included intelligence, psychomotor skill, mechanical comprehension, and spatial measures. Figure 3 shows an example of a World War II vintage psychomotor test used by the US Army Air Corps. The British (Parry, 1947) and the Canadians (Signori, 1949) employed tests similar to those of the Americans. Hilton and Dolgin (1991) recorded that the Germans used ability measures similar to those used by the Allies. Geldard and Harris (1946) investigated the pilot selection system used by the Japanese during World War II and found that they were using tests based on the American Army Alpha, a paper-and-pencil derivative of the Binet intelligence test.

The quarter century following World War II showed little change in pilot selection methods. The United States, like most countries, was limited to producing new forms of paper-and-pencil multiple aptitude tests with some exceptions (see for example, Gopher & Kahneman, 1971). The field of personality measurement saw most of the research innovation (see for example Dolgin & Gibb, 1989). Since 1970, studies of multiple aptitudes and psychomotor abilities (Carretta, 1990; Imhoff & Levine, 1981) have been prevalent. For a more complete review, see Hilton and Dolgin (1991) and Hunter (1989).

Recent Validation Studies

Pilot selection procedures used in NATO-member countries vary in content, focus, and method of administration. However, all NATO-member countries employ some form of psychometric testing as part of military pilot selection (Burke, 1993). Psychometric testing involves the measurement of mental traits, abilities, and processes. Examples of psychometric testing approaches include aptitude, simulation-based, and personality tests. Hilton and Dolgin (1991) and Li (1993) provide detailed reviews.

We have previously defined aptitude as "innate or acquired capabilities." Common examples of aptitude tests include traditional paper-and-pencil tests (Bartolo Ribeiro, 1992; Carretta & Ree, 1995a; Martinussen & Torjussen, 1998; Orgaz & Loro, 1993; Prieto et al., 1996), psychomotor tests (Bartolo Ribeiro, 1992; Bartolo Ribeiro, Martins, Vicoso, Carpinteiro, & Estrela, 1992; Burke, Hobson, & Linsky, 1997; Carretta & Ree, 1993, 1994; Gibb & Dolgin, 1989), and computer-based tests (Bailey & Woodhead, 1996; Boer, 1992; Carretta & Ree, 1993).

Methods such as performance tests (Boer, Harsveld, & Hermans, 1997; Delaney, 1992) and simulation-based tests of pilot work samples (Gress & Willkomm, 1996; Spinner, 1991) are less common. Finally, the emphasis placed on personality assessment varies widely in military pilot selection (Burke, 1993; Dolgin & Gibb, 1989).

Aptitude tests. In a recent review of Royal Air Force (RAF) aircrew selection methods, Bailey and Woodhead (1996) stated that historically the RAF has relied heavily on ability for job specialties such as pilot and on measures of personality/character and biographical information for overall officer suitability. The RAF takes a "domain-centered" approach to test battery

construction. The emphasis is on first identifying the appropriate ability domains for a particular occupation (e.g., pilot, navigator, weapons director, air traffic controller), then choosing one or more tests to represent the critical domains. As the result of task analyses, the RAF identified five aircrew-ability domains: attention capacity (AC), mental speed (MS), psychomotor (PM), reasoning (R), and spatial (SP). The current RAF Pilot Aptitude composite samples all five domains: Control Velocity (PM: anticipatory tracking), Sensory Motor Apparatus (PM: compensatory tracking), Instrument Comprehension (R and MS: interpretation of aircraft instruments, reasoning, and mental speed), Vigilance (AC: monitoring and attention), and Digit Recall (AC: short-term memory). Bailey and Woodhead report the predictive validity of the Pilot Aptitude composite against Basic Flying Training[i] outcome as $\underline{r}$ = .52 after correction for statistical artifacts. This value is considered good and is consistent with general personnel selection results (Schmidt & Hunter, 1998). The RAF computer-based test system is commercially available and has been purchased by several civilian airlines and military services. As of 1997, it was being used for all military pilot selection in the United Kingdom (Burke et al., 1997).

Burke et al. (1997) reported a meta-analysis of the validity of three tests from the RAF pilot selection battery. Meta-analysis techniques allow researchers to combine validity estimates from multiple studies and correct for the effects of statistical and measurement artifacts (Hunter & Schmidt, 1990). Meta-analytic studies are valuable because they provide more accurate estimates of validity than do individual studies. Burke et al. examined the validity of the RAF Control of Velocity, Instrument Comprehension, and Sensory Motor Apparatus tests and a summed composite of the three tests. The sample consisted of 1,760 pilot trainees (RAF fixed-wing, n = 849; Turkish Air Force fixed-wing, n = 570; and British rotary-wing, n = 341). The criterion was a dichotomous pass/fail training outcome score. Observed validities were corrected for range restriction (Thorndike, 1949) and dichotomization (Cohen, 1983) as suggested by Hunter and Schmidt (1990). The validities for the three tests ranged from .15 to .16 in observed form and were from .28 and .29 after correction for range restriction and dichotomization of the criterion. The observed validity of the composite was .24 and increased to .40 after correction. These are reasonably good values considering the lack of perfect reliability of the predictor and criterion variables.

Further, support for the utility of various aptitude tests has been provided in three recent meta-analyses (Hunter & Burke, 1994; Martinussen, 1996; Martinussen & Torjussen, 1998). Hunter and Burke (1994) conducted a "bare bones" analysis of validities for 68 pilot selection studies published between 1940 and 1990. A bare bones analysis corrects for sampling error, but usually does not correct for other study artifacts such as reliability and range restriction. In general, bare bones analyses are less informative than studies fully corrected for artifacts (Hunter & Schmidt, 1990). Mean validities were estimated for 16 predictor categories: general ability, verbal, quantitative, spatial, mechanical, general information, aviation knowledge, gross dexterity, fine dexterity, perceptual speed, reaction time, biodata inventory, age, education, job sample, and personality. All categories except age and personality are examples of aptitude. Biodata inventories may include indicators of aptitude (e.g., education or job-related experience), attitudes, personality, life experiences, etc. Cumulative sample sizes varied by category and ranged from 2,792 (fine dexterity) to 52,153 (spatial). The predictor categories with the highest observed mean validities were job sample (.34), gross dexterity (.32), mechanical (.29), and

reaction time (.28). The predictor groups with the lowest observed mean validities were education (.06), age (-.10), fine dexterity (.10), and personality (.10). However, because these results represent a bare-bones meta-analysis, the validities represent conservative estimates. For instance, measures of general ability, which had a mean observed validity of .13 in the Hunter and Burke meta-analysis, have shown much greater validity when corrected for statistical artifacts such as range restriction and reliability (Ree & Carretta, 1996, 1997).

Martinussen (1996) conducted a meta-analysis of 50 studies published between 1919 and 1993. Most of the studies were published after World War II with 1973 as the median year of publication. The studies were from 11 different countries with about half from the United States and about 74% from English-speaking countries. Martinussen grouped the predictors into nine categories. Aptitude categories included cognitive, intelligence, psychomotor/information processing, aviation information, combined index (a combination of several tests, usually cognitive and psychomotor), academics, and flying training experience. Non-aptitude categories included personality and biographical. Cumulative sample sizes for the predictor categories ranged from 3,736 (aviation information) to 17,900 (cognitive). Although Martinussen corrected the validities for dichotomization of the criterion, as with Hunter and Burke (1994), due to a lack of data, correlations were not corrected for range restriction or for reliability of the criterion. Martinussen's validities were similar to those reported by Hunter and Burke. The highest validities were for the combined index (.37) and flying training experience (.30). The next highest validities were for cognitive (.24), psychomotor (.24), aviation information (.24), and biographical (.23). The categories with the lowest validities were intelligence (.16), academics (.15), and personality (.14). Again, these validities should be considered conservative estimates since they were not corrected for either range restriction or unreliability. The effects of range restriction on the validity of measures of intelligence and conscientiousness (personality) are particularly notable, as these constructs are pervasive in personnel selection contexts (Schmidt & Hunter, 1998). Pilot training applicants typically have completed a college degree and have been directly selected for training based on academic achievement (e.g., college grades, major) and indicators of conscientiousness (e.g., life experiences, college graduation, direct observation).

Finally, a small-scale meta-analysis was conducted by Martinussen and Torjussen (1998) who examined the validity of Norwegian pilot selection tests from five studies published between 1955 and 1998. Analyses were performed at the test-level and sample sizes ranged from 244 to 977. Validities were corrected for dichotomization of the criterion, but were not corrected for range restriction, unreliability of the criterion, or other statistical artifacts. The highest validities occurred for three tests of technical or job knowledge related to pilot training (see Olea & Ree, 1994): Instrument Comprehension (.29), Mechanical (.23), and Aviation Information (.22). Raven's Matrices (Raven, 1966), often cited as a good measure of general cognitive ability (Jensen, 1998), had a validity of .16. It is interesting to note that no measures of flying experience were included. This may have been a contributing factor in the performance of the pilot technical or job knowledge tests.

What conclusions can be drawn from these meta-analyses? As we have noted, clear interpretation is difficult because in most instances the validities were not corrected for factors that would reduce their magnitudes (e.g., range restriction, unreliability). Despite this limitation, measures of technical/job knowledge, pilot work samples, and flying experience demonstrated

validity against pilot training performance. Measures of psychomotor ability also demonstrated validity. Measures of cognitive ability and personality were less valid. However, as previously noted, this is not surprising as measures of cognitive ability (Carretta & Ree, 1996b; Hunter & Burke, 1998, Ree & Carretta, 1996) and conscientiousness (Weeks & Zelenski, 1998) are mainstays in military pilot selection procedures, thus leading to restriction of range on these constructs.

Simulation-based tests. Although the use of multiple aptitude tests is common in pilot selection, others have proposed using simulator-based tests to improve selection procedures (Gress & Willkomm, 1996; Long & Varney, 1975; Spinner, 1991). Simulator-based tests have an "intuitive appeal" as they look like some part of the job (e.g., instrument flight) for which applicants are being selected.

The US Air Force (Long & Varney, 1975) examined the utility of a computer-based apparatus approach for pilot selection known as the Automated Pilot Aptitude Measurement System (APAMS). The APAMS system was very crude by today's standards (see Ree & Carretta, 1998 for a description). It could accommodate only one participant and test presentation/response materials and was limited in motion to pitch, roll, and yaw. The APAMS syllabus consisted of a 5-hour sample of flying tasks based on the USAF syllabus for the T-41 light aircraft (single engine high-wing monoplane) training program. The APAMS tasks were intended to reflect individual differences in basic psychomotor abilities, learning rates, multi-task integration, and performance under overload. They were not intended to train pilot skills. Participants in the validation study were 178 student pilots. Pilot training criteria from light aircraft (T-41) training were four dichotomous outcome variables based on flying grades (e.g., students rated as "waivered" or "excellent" versus students rated as "good," "fair," or "deficient"). Pilot training criteria for primary (T-37) jet training (twin turbo fan two-seat jet) were two dichotomous variables (graduates versus all eliminees and graduates versus only those eliminated for "flight training deficiencies"). The average multiple correlations for APAMS scores with training performance were .49 for the T-41 criteria and .30 for the T-37 criteria. Despite good predictive validity for training performance, the USAF decided not to pursue full-scale development and operational implementation of the APAMS system due to its cost and poor utility for decentralized testing. However, it should be noted that the Canadian Air Force, which uses centralized testing for pilot selection, has developed and operationally implemented a system known as the Canadian Automated Pilot Selection System (CAPSS; Okros, Spinner, & James, 1991; Spinner, 1991) which is largely based on APAMS.

CAPSS (see Figure 4) is a moving-based simulator of a single-engine light aircraft (Okros et al., 1991; Spinner, 1991). The test system records up to 250,000 instrument readings per candidate. CAPSS testing includes five, 1-hour "flights" that are performed over a 2.5-day period. In addition to test administration time, participants must spend time preparing for each test session (i.e., reviewing instructions and a flight plan). Over the five test sessions, participants are instructed on, practice, receive feedback, and perform eight basic flight maneuvers. Spinner (1991) examined the validity of CAPSS scores for predicting completion (pass/fail) of preliminary flying training (PFT) for 172 participants. PFT consists of classroom instruction and 27 hours on a CT-134 Musketeer. Spinner reported a multiple correlation of .47. Using discriminant analysis[ii], Okros et al. (1991) subsequently examined the utility of CAPSS scores

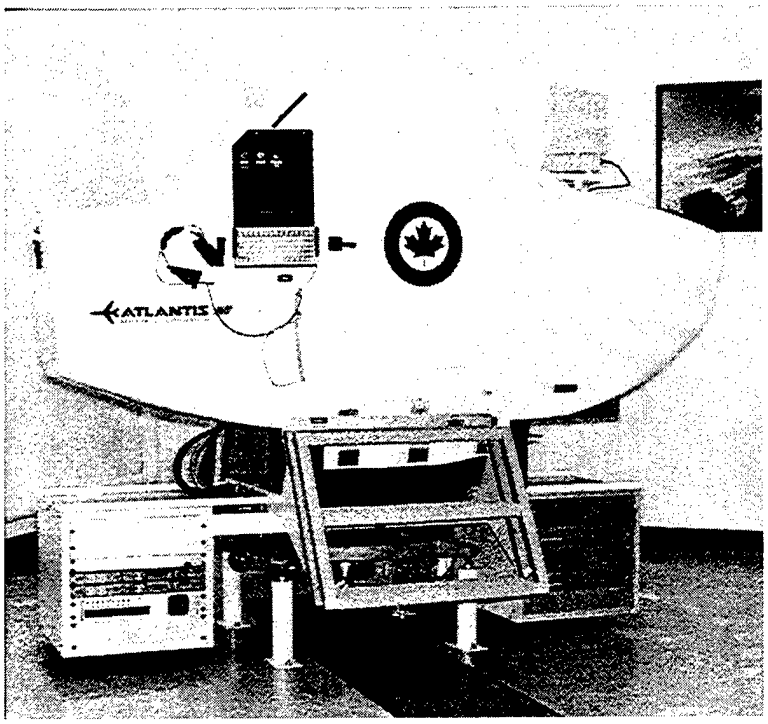


Figure 4. Canadian Automated Pilot Selection System (CAPSS)

for identifying graduates and failures in both PFT (using the Spinner, 1991 sample) and basic flying training (BFT). BFT is the initial jet-training course. CAPSS scores correctly classified 75% (129 of 172) of the pilot trainees attending PFT. For BFT, CAPSS scores correctly classified 80% (154 of 192) of the pilot trainees.

Another example of a simulator-based pilot aptitude test is the FPS 80 (Gress & Willkomm, 1996). The FPS 80 is used as part of a multi-step, sequential selection strategy that includes officer selection, basic tests (psychological and medical), officer school, FPS 80, and flight screening. Successful candidates go on to jet, transport, or helicopter training.

The FPS 80 is a low-fidelity simulator of a single-engine propeller-driven aircraft that consists of a control center and two cockpits. The flight model is based on the Piaggio 149D (single engine low-wing monoplane). Pilot candidates complete four "missions" on the FPS 80 over a two-week period. Prior to performing the missions, pilot candidates are given a training guide that includes detailed descriptions of the missions to be flown. They also must complete two lessons on basic aerodynamic principles. Prior to the first mission, candidates must pass a written test on mission-relevant material. FPS 80 performance is graded in two ways: a computer-generated score based on data from the check ride and an observation-based rating by an aviation psychologist. The psychologist rates each candidate on several factors (e.g., aggressiveness, concentration, coordination, stress tolerance, training progress). A single composite score is generated across all four missions that combines the computer-generated and observation-based scores. Gress and Willkomm (1996) evaluated the validity of the FPS 80 and the basic psychological selection tests for student pilots attending flight screening. The criteria

consisted of academic grades (n = 310) and a final flying score (n = 267) during flight screening. Results indicated that the basic psychological tests had uncorrected validities of .24 and .30 against academic grades and final flying score, respectively. Using the FPS 80 grade along with the basic psychological tests increased the validities to .42 and .54, values that are consistent with meta-analytic findings (Schmidt & Hunter, 1998). Although Gress and Willkomm were encouraged by the results of the validation study, they identified several obstacles to the use of simulator-based tests for pilot selection including cost of the test system and test administration (e.g. centralized testing, amount of time needed).

To summarize, the validity of simulation-based approaches for pilot selection appears comparable to that for general cognitive ability. Further, simulation-based tests may significantly increment the validity of cognitive tests when the two approaches are used together (Gress & Willkomm, 1996). These results are consistent with a large-scale meta-analysis of 19 commonly used personnel selection methods across many occupations (Schmidt & Hunter, 1998). Schmidt and Hunter reported meta-analytically-derived validities of .51 for *g* and .54 for work sample tests for predicting job performance. When used together, the multiple correlation was .63.

Despite their apparent validity and incremental validity, simulator-based tests have drawbacks involving the costs associated with test development and administration (centralized testing, single-administration, preparation and administration time). These drawbacks make simulation-based tests impractical for evaluating large numbers of applicants or for pilot selection programs that rely solely on decentralized testing. Simulation-based testing probably has its greatest value in multiple-stage selection situations where applicants first could be screened on inexpensive group-administered paper-and-pencil cognitive tests. Those who "passed" this screen could be brought to a centralized location for simulator-based testing. We are not aware of any studies done to determine the cost-benefit tradeoffs of simulator-based tests for pilot selection. To be useful, the cost of test development and administration would need to be made up by a reduction in training costs (e.g., reduced attrition, reduced training requirements).

Personality. The relation between personality factors and military pilot performance has been the subject of many studies (see Dolgin & Gibb, 1989 for a descriptive review). NATO-member countries vary substantially in the emphasis placed on personality assessment in pilot selection, as well as on assessment methods (Burke, 1993). Some countries, such as the United Kingdom and the United States do not directly measure personality during selection. In these cases, personality assessment may find its way into the selection process indirectly through its influence on training commander's ratings of cadets on "officership" and "military bearing" or through interviews and observer ratings. The range of explicit personality measures that have been evaluated and are in use within NATO includes a variety of paper-and-pencil questionnaires, projective tests, clinical interviews, and computer-based measures that appear to combine ability and personality assessment. Some examples include the Eysenck Personality Inventory (Jessup & Jessup, 1966), Jackson's Personality Research Form (Retzlaff & Gilbertini, 1987), and the Defense Mechanisms Test (Harsveld, 1991; Martinussen & Torjussen, 1993). Often, personality tests are used during an interview with a psychologist (e.g., Evdokimov, 1988).

Despite the large number of personality characteristics examined using a variety of instruments and methods, empirical support regarding the role of personality in pilot performance is lacking (Dolgin & Gibb, 1989). It is not unusual to find studies that have used the same personality instrument to predict pilot performance yielding contradictory results. An illustrative example is provided by the Defense Mechanism Test (DMT), a projective test in which the participant is shown a picture for a short exposure using a tachistoscope. The participant is asked to describe or draw their impression of the images shown. The responses are scored according to Freudian defense mechanisms. While validations reported for Scandinavian researchers yield impressive results (Torjussen & Vaemess, 1991), a British study of RNAF pilots found zero validity against flying training attrition (Harsveld, 1991). Burke (1993) speculated that cultural factors might have caused these contradictory results for the DMT.

Developments in personality theory in the late 1980's indicated that past reviews of the personality-performance literature suffered from a lack of a conceptual framework for evaluating results from different studies. A consensus model of personality emerged, based on the observation that five global factors adequately describe individual differences in personality traits (Digman, 1990; Tupes & Christal, 1961). These factors, which are known as the "Big Five," are Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness. Each of these broad factors includes several facets. For example, Conscientiousness includes the facets of Achievement Striving, Competence, Dutifulness, Deliberation, Order, and Self-Discipline. The utility of the Big Five framework has been demonstrated in several meta-analytic studies of the relations between personality and job performance (Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991). In another meta-analytic study, Schmidt and Hunter (1998) demonstrated that measures of Conscientiousness are incrementally valid for predicting training and job performance when paired with measures of general cognitive ability.

The study of personality in military pilots generally follows two lines. In one, personality profiles of pilot applicants, trainees, or pilots are compared to the general population (e.g., Callister, King, Retzlaff, & Marsh, 1999). In the other, personality scores are validated against some indicator of training or job performance (Dolgin & Gibb, 1989). Callister et al. (1999) used the Revised NEO-PI (Costa & McCrae, 1992), a measure of the Big Five, to develop personality profiles for male and female USAF student pilots. Participants were 1,301 (1,198 males and 103 females) student pilots tested during a flight screening program. Compared with the general population, student pilots scored high on Extraversion (83^{rd} percentile), Openness (60^{th} percentile), and Conscientiousness (58^{th} percentile), and low on Neuroticism (42^{nd} percentile) and Agreeableness (20^{th} percentile).

Siem and Murray (1994) used the Big Five framework to investigate personality factors affecting pilot combat performance. Participants were 100 USAF pilots. Most (90%) were Captains with a minimum of six years service. Several (43%) had combat experience in Operation Desert Storm. Participants rated the importance of 60 personality traits for each of six flying performance dimensions. The 60 traits were selected from unipolar markers of the Big Five developed by Goldberg (1992). The six flying performance dimensions were 1) flying skills and knowledge, 2) compliance, 3) crew management and emotional support, 4) leadership, 5) situational awareness, and 6) planning. Conscientiousness was rated as the most important factor

for five of the six performance criteria. Openness was rated slightly higher than Conscientiousness for the planning dimension. No predictive validation study was done.

As previously noted, the utility of cognitive and psychomotor abilities for predicting military pilot training performance has been illuminated from the application of meta-analytic techniques (e.g., Hunter & Burke, 1994; Matinussen, 1996; Martinussen & Torjussen, 1998). We believe that the role of personality factors in pilot performance would benefit from the joint application of the Big Five framework and meta-analysis.

Current Research

The previous section described several approaches to pilot selection. This section concentrates on examining what underlying constructs are measured by pilot selection tests and what about them is predictive of training and job performance. We chose to focus on USAF research, as the USAF has been a leader in this area.

Our discussion is guided by the seminal work of Schmidt and Hunter (1998) who examined the validity of general mental ability (*g*) and 18 other selection procedures for predicting training and job performance across many occupations. On the basis of meta-analytic findings, Schmidt and Hunter concluded that the three combinations of predictors with the highest validity and utility for job performance were *g* plus a work sample test, *g* plus an integrity test (or a conscientiousness test), and *g* plus a structured interview. They also concluded that the latter two methods were appropriate for both entry-level selection and selection of experienced employees. It should be noted that Schmidt and Hunter did not consider measures of job knowledge or work sample performance for entry-level jobs, as these methods were not commonly used for that purpose. However, as we have already discussed, in pilot selection job knowledge and work sample tests are fairly common for entry into *ab initio* flying training programs.

The construct of general cognitive ability, *g*, was developed in the first decade of the 20th century by Charles Spearman. Every test or measure of ability measures *g* and specific ability or knowledge, *s*. A long history of research findings has demonstrated *g* to be the most valid predictor of academic performance, job performance, and for numerous other human characteristics (Brand, 1987; Jensen, 1998; Schmidt & Hunter, 1998). The predictive validity of *s* is mostly due to specific knowledge, not specific ability. General cognitive ability is usually defined as the common source of variability among a set of cognitive measures. For practical purposes, it can be thought of as the main factor of intelligence. Jensen (1998) provides the most complete presentation and discussion of general cognitive ability.

Despite these consistent results, the subject of *g* remains contentious (McClelland, 1993; Ree & Earles, 1992, 1993; Schmidt & Hunter, 1993; Sternberg & Wagner, 1993). Some have proposed that to understand human characteristics and job performance, it is necessary to measure noncognitive traits, specific abilities, and knowledge. For example, McClelland (1993) suggested that under some circumstances noncognitive traits such as motivation might be better predictors of job performance than cognitive abilities. Sternberg and Wagner (1993) proposed using measures of tacit knowledge and practical intelligence instead of measures of "academic

Table 1.
Composition of AFOQT Aptitude Composites

Test/Abbr.	Composite V	Q	AA	P	N-T	Description
Verbal Analogies (VA)	X		X	X		Ability to reason & recognize word relationships
Arithmetic Reasoning (AR)		X	X		X	Understanding of arithmetic relationships expressed as word problems
Reading Comp. (RC)	X		X			Reading skill
Data Interpretation (DI)		X	X		X	Ability to extract data from graphs & charts
Word Knowledge (WK)	X		X			Understanding of written language through the use of synonyms
Math Knowledge (MK)		X	X		X	Use of mathematical terms, formulas, & relationships
Mechanical Comp. (MC)				X	X	Understanding of mechanical functions
Electrical Maze (EM)				X	X	Spatial ability based on choice of a path through a maze
Scale Reading (SR)				X	X	Ability to read dials & scales
Instrument Comp. (IC)				X		Ability to determine aircraft attitude from illustrations of flight instruments
Block Counting (BC)				X	X	Spatial ability through analysis of 3-dimensional representations of blocks
Table Reading (TR)				X	X	Ability to quickly & accurately extract information from tables
Aviation Information (AI)				X		Knowledge of general aviation technology & concepts
Rotated Blocks (RB)					X	Spatial aptitude through mental manipulation & rotation of objects
General Science (GS)					X	Knowledge of scientific terms, concepts, & principles
Hidden Figures (HF)					X	Spatial ability to find simple figures embedded in complex drawings

intelligence." They define tacit knowledge as "the practical know how one needs for success on the job" (p. 2). Practical intelligence is defined as a more general form of tacit knowledge. Schmidt and Hunter (1993) noted that Sternberg and Wagner's concepts of tacit knowledge and practical intelligence are redundant with the well-established construct of job knowledge.

Currently, the Air Force Officer Qualifying Test (AFOQT; Carretta & Ree, 1996a; Skinner & Ree, 1987) is an important component in USAF pilot selection. As shown in Table 1, it consists of 16 cognitive and pilot job knowledge tests. The tests are combined into three academic composites used primarily to assess "officership" (Verbal, Quantitative, and Academic Aptitude) and two aviation-related composites (Pilot and Navigator-Technical). The Pilot composite is different from the other four AFOQT composites in that it includes tests of job knowledge (i.e., Instrument Comprehension and Aviation Information tests).

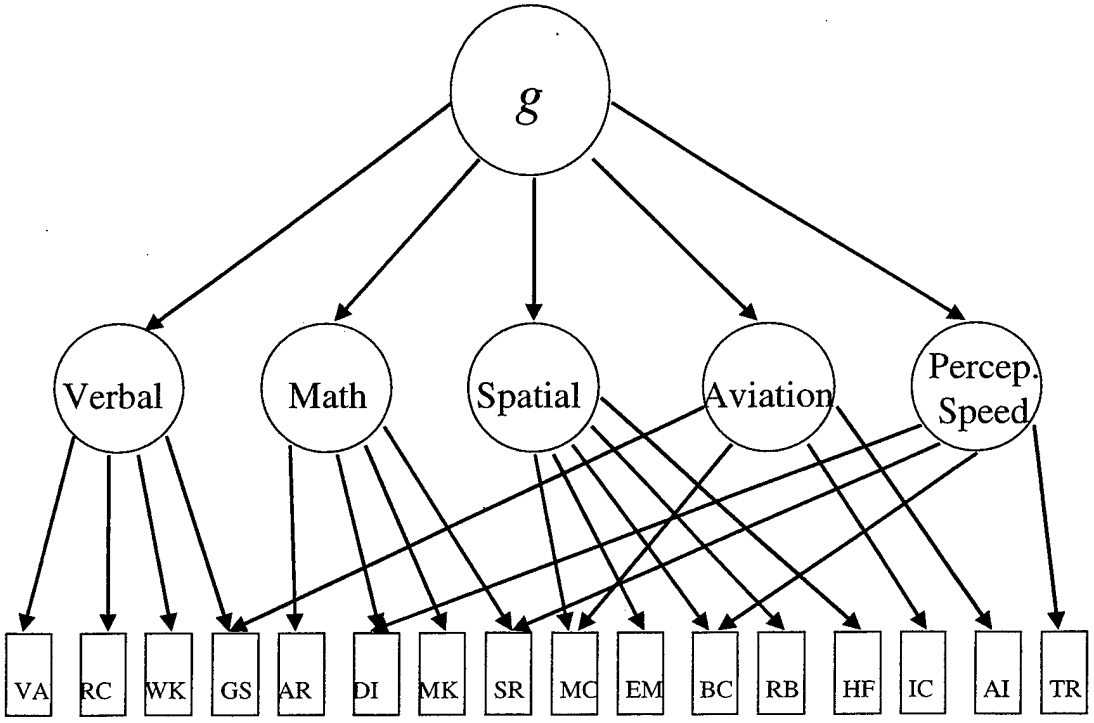


Figure 5. Hierarchical factor structure of the AFOQT with *g* as the higher-order factor and five lower order factors of Verbal, Math, Spatial, Aviation Knowledge, and Perceptual Speed.

Confirmatory factor analysis (CFA; Jöreskog & Sörbom, 1996; Kim & Mueller, 1988) is a statistical technique that allows investigators to specify and test hypotheses about the relations among a set of variables. Recent CFAs (Carretta & Ree, 1996a) have found that the AFOQT displays a hierarchical nature similar to other multiple aptitude test batteries (Jensen, 1994; Ree & Carretta, 1994a; Vernon, 1969). See Figure 5. The higher-order factor was identified as general cognitive ability (*g*). All 16 tests contributed to the measurement of *g*. The proportion of common variance due to *g* was 67%. The remaining common variance (33%) in the residualized (Schmid & Leiman, 1957) lower-order factors was 11%[iii] for verbal, 9% for aviation

interest/aptitude, 4% for perceptual speed, 4% for spatial, and 4% for math. These proportions are similar to that found in other multiple aptitude batteries (Jensen, 1980). Most of the predictive utility of the AFOQT against pilot training performance can be attributed to its measurement of *g* and aviation interest/aptitude (Olea & Ree, 1994; Ree, Carretta, & Teachout, 1995).

Another important component in USAF pilot selection is the computer-administered Basic Attributes Test (BAT; Carretta, 1992a). The BAT test system (see Figure 6) is fairly representative of computer-based pilot aptitude tests. The test apparatus consists of a computer and monitor built into a testing carrel. The carrel has side, back, and top panels designed to minimize glare and distractions. Participants respond to the test stimuli by manipulating individually or in combination, a dual-axis right-hand control stick, a single-axis left-hand control stick, and a specialized response keypad. The BAT battery provides psychomotor test scores (multi-limb coordination, pursuit tracking, rate control, response time), cognitive scores (short-term memory), and a personality measure (attitudes toward risk).

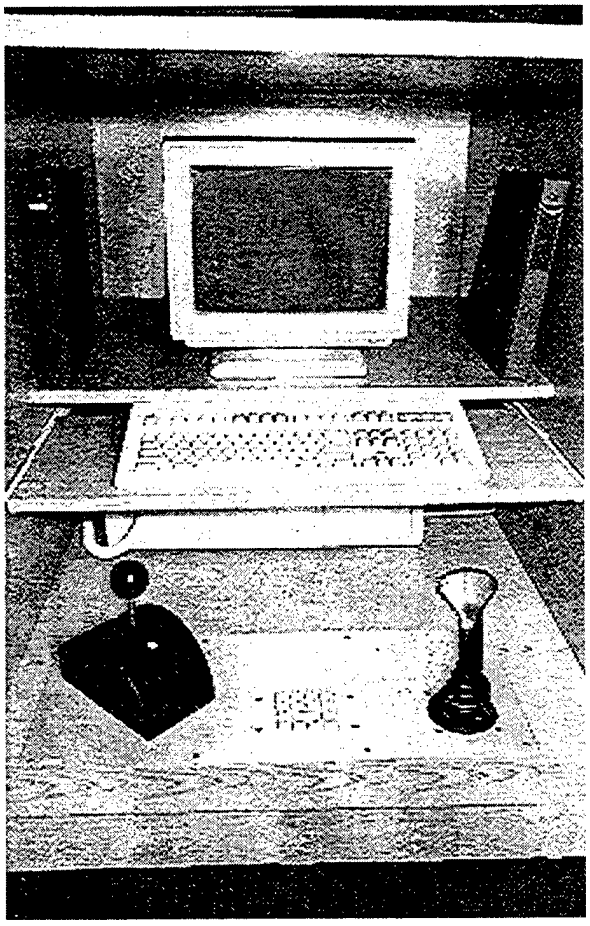


Figure 6. Basic Attributes Test (BAT) System - a computer-based test system currently used in US Air Force pilot selection.

The US Air Force combines the AFOQT Pilot composite, BAT psychomotor (multi-limb coordination, control precision, reaction time, and rate control), cognitive (short-term memory), and personality (attitudes toward risk) measures, and a self report of flying experience to create a measure of pilot aptitude known as the Pilot Candidate Selection Method (PCSM; Carretta, 1992a). The BAT psychomotor scores are used in a unit-weighted composite as are the tests in the AFOQT Pilot composite. These psychomotor and AFOQT Pilot composites are regression-weighted along with the other PCSM components of cognitive, personality, and flying experience to predict pilot training criteria. Several studies have demonstrated the validity of PCSM scores for a variety of pilot training criteria. High PCSM scores are associated with greater probability of successfully completing jet training (Carretta 1992a, 1992b, 2000), fewer flying hours needed to complete training (Duke & Ree, 1996), higher class ranking (Carretta, 1992b), and greater likelihood of being fighter-qualified (Weeks, Zelenski, & Carretta, 1996).

Carretta and Ree (1994) evaluated the validity and incremental validity of the components of a pre-operational form of PCSM on a sample of 678 pilot trainees. Analyses showed the following (uncorrected for range restriction or dichotomization) correlations with passing-failing pilot training: AFOQT Pilot composite .17, BAT psychomotor .15, BAT cognitive (information processing) .06, BAT personality .10, and flying experience .17. Adding the BAT and flying experience scores to the AFOQT Pilot composite raised the correlation from .17 to .30. It should be noted that the correlation magnitudes were reduced due to the severe dichotomization of the criterion produced by the high pass rate in this sample (85.7% passing). For this chapter, we corrected the correlations for dichotomization of the criterion. As expected, all correlations increased after correction for dichotomization. The corrected correlations were AFOQT Pilot composite .26, BAT psychomotor scores .23, information processing .09, personality .16, and flying experience .26. When the AFOQT Pilot, BAT, and flying experience scores were used together, the multiple correlation was .46. These correlations, although good, should be considered conservative estimates, as they were not corrected for range restriction.

Incremental validity of the BAT and flying experience scores relative to the AFOQT tests was estimated using correlations corrected for range restriction and for dichotomization of the passing/failing criterion. The AFOQT tests were the best predictors of pilot training performance with a multiple correlation of .42. Other studies have shown that the predictive validity of the AFOQT for pilot training comes mostly from its measurement of *g* and pilot job knowledge (Olea & Ree, 1994; Ree et al., 1995). Using the AFOQT tests and adding the BAT and flying experience scores, the multiple correlation increased to .52. This 24% increase and increment of .10 above the AFOQT tests represents potentially large cost avoidance savings for the US Air Force. Cost estimates of each person who fails to complete USAF undergraduate pilot training range from $50,000 (Hunter, 1989) to $80,000 (Siem et al., 1988). These should be considered conservative estimates, as they are over a decade old. Obviously, even a small reduction in training attrition due to improved selection could produce large training cost avoidance savings.

Studies were conducted to determine the causes of the level of the incremental validity of the various predictors. The BAT psychomotor scores were investigated in the presence of a highly *g*-loaded battery of verbal and mathematical cognitive tests (Ree & Carretta, 1994b). Confirmatory factor analyses on a sample of 354 enlisted personnel showed the BAT psychomotor scores to have three lower-order factors: Two-Hand Coordination, Complex

Coordination, and Time Sharing. As expected, the cognitive battery showed two lower-order factors of verbal and math. A higher-order psychomotor factor influencing all psychomotor scores was found and unexpectedly, a higher-order *g* was found to influence all scores, both cognitive and psychomotor. Subsequently, Wheeler and Ree (1997) examined the validity of measures of general and specific psychomotor abilities extracted from the BAT psychomotor tests for predicting USAF pilot training performance. Results indicated that the validity of the BAT psychomotor tests comes from their measurement of a general psychomotor factor and *g*.

The choice of the personality trait of attitude toward risk in the PCSM equation was made before the current prevalence of the Big Five (Digman, 1990) model of personality. Our finding of small incremental validity for personality variables is similar to the findings of McHenry, Hough, Toquam, Hanson, and Ashworth (1990).

That information processing speed was not predictive was surprising (Carretta & Ree, 1994) as it has been found to be predictive of pilot training achievement in other studies (Carretta, 1992a). We suspect that the lack of validity in this study may have been a consequence of sampling error.

Consider another example. Olea and Ree (1994) compared the validity of general cognitive ability, *g*, and specific abilities (including pilot job knowledge), $s_1 \ldots s_n$ for predicting several pilot criteria in samples ranging from 1,867 to 3,942. General cognitive ability and specific abilities (including pilot job knowledge) were estimated from the AFOQT. The criteria included academic performance and work samples of landings, loops, and rolls. There was also an overall performance composite. The two most notable criteria were the overall performance composite and the work samples. The overall performance composite was a sum of the individual criteria providing a global measure of training performance. The work sample criteria were more like job performance measures of core technical task proficiency (i.e., core content specific to the job) and general task proficiency (i.e., general or common tasks not job-specific), such as used by McHenry et al. (1990). Multiple correlations were compared to estimate the predictive efficiency of *g* and *s* for each of the criteria. Notwithstanding the apparent differences among the criteria, *g* was the best predictor while s contributed little. The validity for *g* ranged from .21 to .43 across all pilot criteria with a mean of .31. The incremental validity for the specific abilities beyond *g* ranged from .07 to .14 with a mean of .10. Little incremental validity was found for the composite performance criteria for pilots (.09) or for the work sample criteria. For pilots, three predictors entered each equation: *g*, $s1_2$ and *s3*. Although the exact psychological nature of *s1* and *s3* cannot be assessed with certainty, the weights associated with the components emphasized special knowledge of aviation information and instrument comprehension. Results suggested that the incremental validity of specific measures for pilots was due to specific knowledge about aviation principles, aviation instruments, and aircraft controls rather than specific abilities such as spatial or perceptual ability.

Research results point to *g* as the most important underlying construct in the prediction of pilot training success. Clearly, three others have been shown to be important but to a smaller degree: flying job knowledge, personality, and general psychomotor ability.

Group Differences in Pilot Selection and Training

The study of sex and ethnic group differences is an important consideration in personnel measurement and selection. Several principles must be considered when addressing the measurement of ability in sex or ethnic groups. These include whether the selection instruments measure the same factors for all groups (i.e., factorial invariance), group mean score differences, differential validity, and the causal role of selection factors on the acquisition of job knowledge and skill. McArdle (1996), among others, contends that factorial invariance (i.e., equality of factor loadings) should be demonstrated before other group comparisons (e.g., mean differences, validity) are considered. If factorial invariance is not observed, the psychological constructs being measured may be qualitatively different for the groups being compared, thus clouding the interpretability of other comparisons. For example, how can group mean differences be interpreted if the tests do not measure the same constructs for the groups? (i.e., differential construct validity).

In addition to factorial invariance, it must be shown that the tests do not lead to adverse impact and are not differentially predictive for different groups tested. Adverse impact occurs when members of one group are disproportionally disqualified compared to members of another group on the basis of test performance. Differential prediction would exist if a test were valid for one group of pilot trainees and not valid for some other group of pilot trainees. U. S. Government guidelines discourage the use of personnel selection tests that display adverse impact and prohibit tests that show bias. Evidence of differential prediction is accepted as evidence of test bias.

Factor structure. Group factor structure comparisons have been done for both of the major USAF pilot selection tests (AFOQT and BAT). Factorial invariance was examined by comparing confirmatory factor analytic models for sex and race/ethnic groups. Factorial invariance is established when the factor loadings are the same for the groups being compared (Alwin & Jackson, 1981; McArdle, 1996). A χ^2 test was conducted to determine if the loadings for a score on a factor were the same for the groups being compared (Bentler, 1989).

Carretta and Ree (1995b) examined AFOQT factor structure for large samples of USAF officer applicants (219,887 males, 50,081 females, 212,238 Whites, 32,798 Blacks, 12,747 Hispanics, 9,460 Asian-Americans, and 2,551 Native Americans). The model tested had been confirmed by Carretta and Ree (1996a; see Figure 4). The hierarchical factor is *g* and the five lower-order factors are verbal, math, spatial, aviation interest/aptitude, and perceptual speed. Despite group mean score differences on the 16 AFOQT tests, the model showed good fit for both sex groups and for Whites versus each of the other four race/ethnic groups. Further, the proportions of total and common variance accounted for by *g* and the five lower-order factors were similar for all groups. Results indicated nearly identical structure of ability for sex and race/ethnic groups.

Carretta (1997) compared the factor structure of the Basic Attributes Test (BAT) for 4,888 male and 465 female USAF pilot applicants. The model tested was based on results from earlier confirmatory factor analyses of the AFOQT (Carretta & Ree, 1996a) and the BAT psychomotor tests (Ree & Carretta, 1994b). The model included general cognitive (*g*) and

general psychomotor (PM) factors and lower-order factors representing verbal, math, two-hand coordination, complex coordination, response time, time-sharing, and activities interests. All test scores, including those from the BAT, contributed to *g*. The model demonstrated good fit and the proportion of common and total variance accounted for by the factors was similar for both sexes. Again, despite mean score differences on the tests, results indicated near identity of factor structure for men and women.

Mean scores. Comparisons of mean test scores for men and women on pilot aptitude tests have been done for the USAF AFOQT (Carretta, 1997a) and BAT (Carretta, 1997b), and for RAF computer-based tests (Burke, 1995). Race/ethnic group comparisons have been reported for the AFOQT (Carretta, 1997a). In these studies, the size of the mean differences was expressed in standard deviation units or d (Cohen, 1988). The standard deviation for *d* was defined as the within-group standard deviation ($SD = [Sp^2/n_1 + Sp^2/n_2]^{1/2}$, where $S_p^2 = [SS_1 + SS_2]/[n_1 + n_2 - 2]$; see for example, McNemar, 1969, p. 115) calculated from the weighted average of the variances for the samples being compared (e.g., males and females). Thus, $d = (M_1 - M_2)/SD$. Cohen (1988) characterizes a *d* of .20 as small, .50 as medium, and .80 as large. It should be noted, however, that even "small" *d* values can have a large impact on the proportion of applicants in the lower mean group that would meet or exceed some minimum cut score for selection. Group mean differences were tested using one-tailed t-tests (i.e., majority group – minority group).

Carretta (1997a) examined group mean differences for the AFOQT composites and tests for USAF officer applicants and pilot candidates. Male officer applicants (n = 219, 887) significantly ($p < .05$) outperformed females (n = 50,081) on all composites and 15 of 16 tests (Verbal Analogies test was the exception). The mean *d* value for the composites was 0.42 and ranged from 0.69 (Pilot) to 0.08 (Verbal). The mean *d* value for the 16 tests was 0.44 and ranged from 0.02 (Verbal Analogies) to 0.95 (Mechanical Comprehension). Results for pilot trainees were very different (9,239 males and 237 females). The mean *d* value for the composites was –0.10 (slightly favoring women) and ranged from –0.48 (Verbal) to +0.20 (Navigator-Technical). For the 16 tests, the mean difference was 0.08 (slightly favoring men) and ranged from –0.63 (Verbal Analogies) to +0.84 (Mechanical Comprehension).

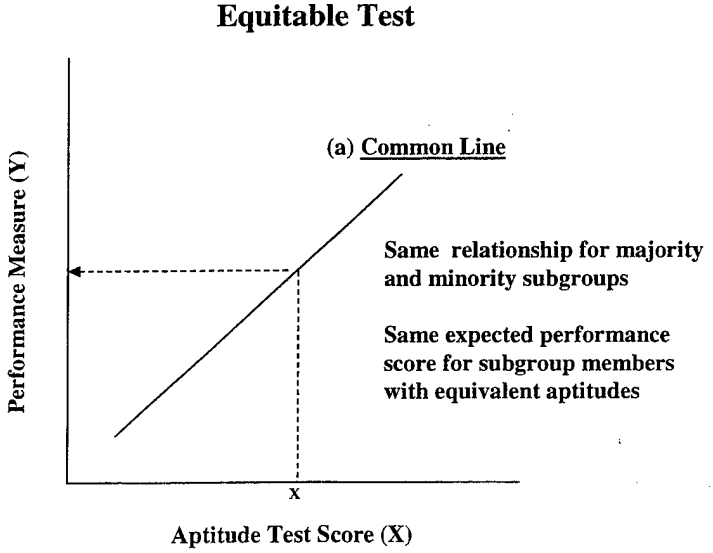
A similar pattern was observed for comparisons of Whites with Blacks and Hispanics (Carretta, 1997a). The officer applicant sample included 212,238 Whites, 32,798 Blacks, and 12,647 Hispanics. The pilot trainee sample included 8,995 Whites, 185 Blacks, and 172 Hispanics. Large mean score differences in the officer applicant sample were reduced in the pilot trainee sample. The average White-Black composite difference was 1.22 *d* for officer applicants and 0.52 *d* for pilot trainees. The average White-Hispanic differences were 0.80 *d* and 0.40 *d* in the officer applicant and pilot samples. The reduction in mean differences between Whites and other groups in the pilot sample was interpreted as a direct result of the selection process.

Carretta (1997b) observed small to large mean score differences for 4,888 male and 465 female USAF pilot applicants tested on the BAT. No race/ethnic group comparisons were reported. All mean score differences favored males and were statistically significant. The smallest difference occurred for a cognitive measure of short-term memory (0.10 *d*) and the largest difference for a psychomotor composite (1.68 *d*).

Burke (1995) examined mean score differences for male and female RAF pilot applicants. Grouping the tests by content, all mean score differences favored males (though no significance levels were reported). The mean *d* values were: information processing 0.02, perceptual speed 0.10, reasoning 0.21, and psychomotor 0.98.

Predictive validity. There are numerous studies of sex and race/ethnic group differences in predictive validity for cognitive tests. Most involve comparisons of racial groups. The cumulative evidence overwhelmingly demonstrates that differential validity is almost nonexistent for cognitive tests (Carretta, 1997a; Jensen, 1980; Roberts & Skinner, 1996).

Carretta (1997a) examined the predictive validity of the AFOQT composites against USAF pilot training passing/failing outcome for several groups. The sample included 9,239 men and 237 women; 8,995 Whites, 186 Blacks, and 172 Hispanics. Examination of differential validity involved the testing of linear models. A "full model" was compared to a "restricted model" that contained a subset of the variables from the full model. An F statistic was used to evaluate the change in predictive efficiency between the full and restricted models using the hierarchical step-down method of Lautenschlager and Mendoza (1986). The starting (full) model (Model 1) for each analysis contained separate estimates for the slopes and intercepts for the two groups (males vs. females, Whites vs. Blacks, Whites vs. Hispanics). The first restricted model (Model 2) removed the separate slope estimates, and the second restricted model (Model 3) removed the separate intercept estimates. First, each AFOQT composite was tested for slope bias. If evidence of slope bias was found, the analysis sequence was terminated. If no slope bias was found, the composite was tested for difference in intercepts. Figure 7 shows illustrations of an unbiased test (Figure 7a), intercept (level) bias (Figure 7b), and slope bias (Figure 7c).



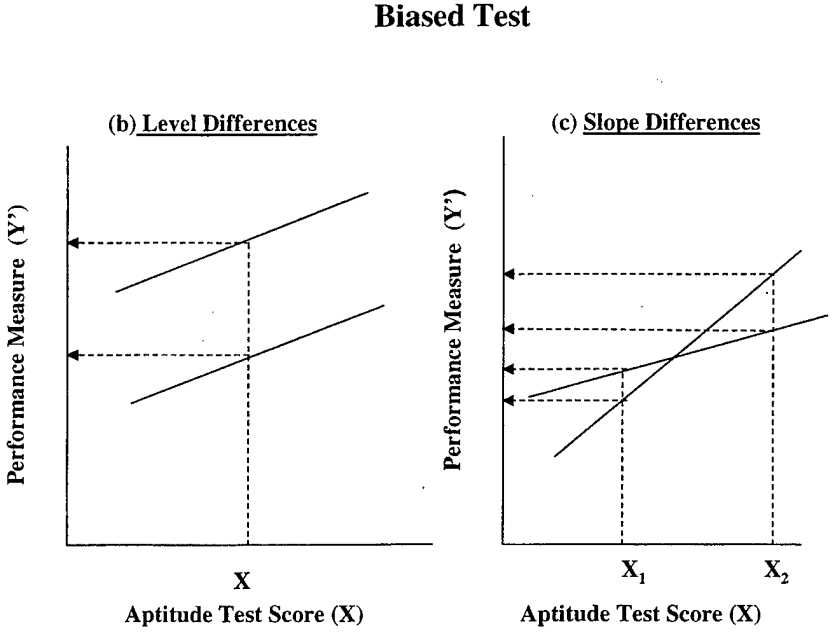


Figure 7. Illustrations of an unbiased test (Figure 7a), intercept (level) bias (Figure 7b), and slope (Figure 7c) bias.

Comparisons of Model 1 versus Model 2 for the AFOQT composites indicated there were no group differences in slopes. Comparisons of Model 2 and Model 3 showed some significant intercept differences. In all instances where intercept differences occurred, performance was overpredicted for the minority group. For the sex group comparisons, passing/failing training was overpredicted for females for the Quantitative and Academic Aptitude composites. For the ethnic group comparisons, no intercept differences were found for Whites versus Blacks. However, passing/failing training was overpredicted for Hispanics relative to Whites for the Pilot, Navigator-Technical, Quantitative, and Academic Aptitude composites. After correction for unreliability of the predictors (Jensen, 1980, p. 384), all differences were reduced to a trivial .0004 or less.

Causal models. Carretta and Ree (1997) tested the Ree et al. (1995) causal model of pilot training (see Figure 1) on separate samples of male (n = 3,369) and female (n = 59) pilots. Figure 8 shows the coefficients for the causal model. The results are considered preliminary due to the small female sample size. Although results were similar for men and women, the direct and indirect influence of *g* on flying performance was stronger for women than for men. Also, the relationship between prior job knowledge and flying performance was stronger for women than for men. Consistent with Ree et al. (1995), the influence of early flying skills on later flying skills was very strong for both sexes.

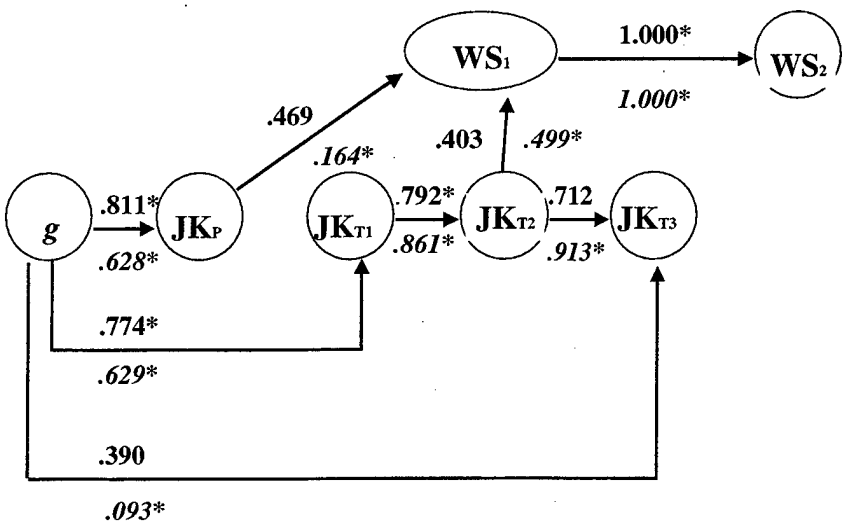


Figure 8. Carretta and Ree (1997) model of the influence of general cognitive ability (*g*) and prior job knowledge (JK_P) on the acquisition of additional job knowledge (JK_{T1} to JK_{T3}) and sequential training performance (WS_1 and WS_2) for male and female pilots. Note. The values for the male sample are in italics.

Summary. Despite group mean score differences on pilot selection tests, confirmatory factor analyses indicated that the same factors were measured for all sex and ethnic groups. In studies of predictive bias, no evidence of differential validity was found for male versus female pilot trainees or for Whites versus racial/ethnic minorities. An examination of causal models of ability and prior flying knowledge on the acquisition of additional flying knowledge and flying skills showed similar structure for men and women.

COMMERCIAL PILOT SELECTION

Almost all of the published literature on pilot selection methods concerns military pilot selection. However, there have been a few recent studies involving commercial aviation. Many commercial airlines rely on "realistic" or high-fidelity simulators because they are hiring trained pilots. This is different from the military's reliance on paper-and-pencil tests and computerized tests for a simple reason. Commercial pilot selection traditionally has relied on the military for trained pilots, a trend that appears to be ending as the military currently is training fewer pilots. Commercial selection procedures will have to become like the military procedures, as the commercial airlines have to select untrained pilot applicants and provide initial training. The following section provides mostly descriptive information collected by survey. Such survey data should be seen as describing the current activities of commercial selection and not as scientific data from which causal explanations should be drawn as to proper practice.

US Air Carriers

In the United States, the Federal Government (Equal Employment Opportunities Commission, 1978) has issued a set of technical standards for validity studies that call for a job analysis to gather information about the job. The executive branch of the Federal Government (i.e., Department of Defense) has exempted itself from these standards. Commercial air carriers are not exempt and they should conduct informative job analyses. McCormick (1976, 1979) and Gael (1988) provide detailed guidelines for conducting job analyses.

The Federal Aviation Administration (FAA) conducted surveys to identify trends in pilot hiring and selection for US air carriers in 1994 (Suarez, Barborek, Nikore, & Hunter, 1994) and again in 1997 (D. R. Hunter, personal communication, August 27, 1998). The surveys focused only on hiring and selection procedures and did not assess the validity of the selection methods.

Suarez et al. (1994) sampled corporate operators, regional/major airlines, specialized air services, and commuters/air taxis. Surveys were conducted by mail and were anonymous. The overall response rate was 20.8% and varied by type of carrier from 12.6% (corporate operators and specialized air services) to 30.3% (commuters and air taxis). As expected, responses to questions about pilot hiring practices and selection methods varied by type of carrier. Across all types of carriers, the reported percentages using the following sources of new hires were: flight school (56%), air taxi (42%), commuters (25%), government/corporate (25%), and major carriers (11%). The most common sources cited by the regional or major carriers were other major carriers (59%), government/corporate (53%), and commuter airlines (53%). Relatively few regional or major airlines reported hiring pilots from either air taxi services (24%) or flight schools (6%).

Carriers reported using a combination of several different selection methods including interviews, aptitude tests, flight checks, simulators, clinical psychological assessment, reference checks, and biographical checks. The most commonly used selection methods across all types of carriers were interviews (96%), reference checks (93%), and flight checks (76%). The least common methods were simulators (17%) and clinical psychological assessment (14%). Among the regional and major carriers, the most common selection methods were reference checks (100%), background checks (100%), and interviews (94%). Simulators (47%), aptitude tests (35%), psychological assessment (25%), and flight checks (24%) were less common. The skills needed to fly an airplane can be checked by a work sample in the form of either a check flight or simulator. Suarez et al. noted that since skill levels can be tested, minimum qualification in combination with prior flying experience are the most widely used set of hiring variables. Smaller operators tend to assess pilot skills using flight checks, while larger operators use simulators. Overall, aptitude tests and psychological assessment were not commonly used selection methods for US air carries perhaps because they are satisfied with indicators of pilot skills (e.g., log books, flight check, simulator) and biographical data (e.g., background and reference checks). To some extent, aptitude and personality factors are assessed in the background and reference checks.

In reviewing reported hiring practices among US carriers, Suarez et al. (1994) concluded "With many pilots available, respondents to this survey appear to be hiring selectively – selecting older pilots with more experience than the minimums required. " (p. 22). However, Suarez et al. speculated that the glut of experienced pilots available to the airlines in 1994 allowing them to be extremely selective in their hiring practices, would not likely endure. Suarez et al. speculated that several factors could lead to an eventual shortage of experienced pilots. These include expected growth in the airline industry, fewer pilots entering the labor pool from the military, reduction in the number of pilots working their way up to the major/regional airlines from air taxi and commuter carriers, and mandatory age-based retirement. Despite the possible shortage of pilots in the future, only 6% of those responding to the survey reported that they intended to change their recruiting and selection procedures to cover the shortfall over the next few years. A 1997 survey of 29 regional and 10 major US air carriers (D. R. Hunter, personal communication, August 27, 1998) reported similar results. When asked to estimate the proportion of new pilots hired by their organizations in 1996 that came from various sources, the most common sources for regional carriers were air taxi operators (23% of new hires), flight instruction schools (20%), and regional airlines (17%). The most common sources of new pilots for the major airlines were the regional airlines (40% of new hires) and the military (37%). As with the 1994 survey, selection interviews, aptitude testing (e.g., flying knowledge, general education/achievement), flying skills evaluations (e.g., review of applicant pilot logbooks, simulator performance) and aircraft check rides were common selection methods. It is interesting to note that there was only a single reference made to use of a personality test. Simulators/training devices were used to assess pilot skills by 67% of the regional carriers and 90% of the major carriers. Twenty four percent (24%) of the regional carriers and 80% of the major carriers reported that the pilot candidate is given a check ride in the type of aircraft they may operate.

Hansen and Oster (1997) identified five pathways for civilian pilots. The first pathway, military pilot training, traditionally has accounted for about 75% of the new hires for major US civilian air carriers. On-the-job training, collegiate training, *ab initio* training and foreign hires account for the rest. Hansen and Oster note that US air carriers have shown reluctance to use either *ab initio* programs or foreign hires. Although *ab initio* training is popular with foreign air carriers (e.g. Lufthansa), US carriers have been unwilling to pay for high-cost *ab initio* programs. This attitude will probably continue as long as applicants from other sources are plentiful. If the supply of former military pilots dwindles, US carriers are likely to turn to on-the-job training and collegiate-based programs to pick up the slack.

Non-US Air Carriers

Several recent studies have described pilot selection procedures for non-US commercial air carriers (Bartram & Baxter, 1996; Doat, 1995; Hörmann & Luo, 1999; Manzey, Hörmann, Osnabrügge, & Goeters, 1990; Novis Soto, 1998; Stahlberg & Hörmann, 1993; Stead, 1991, 1995). However, little information is provided about their validity (Bartram & Baxter, 1996; Hörmann & Luo, 1999; Stahlberg & Hörmann, 1993; Stead, 1991). This important omission makes the studies of little practical or scientific value.

Stead (1995) chronicled changes in Quantas' pilot selection procedures from the 1950's to 1990. Early Quantas pilot selection procedures focused on prior flying experience (with an emphasis on command and multi-engine experience), job knowledge, personality suitability, and medical qualification. These factors were assessed through evaluation of the candidate application form, an interview (usually with the Chief Pilot), and a medical examination. Pilot selection procedures at Quantas changed through the years with a greater emphasis on skill (simulator and flight check) and aptitude testing (ability and personality) and a reduced emphasis on interviews. As described elsewhere (Burke et al., 1997), Stead (1991) examined the validity of several pilot aptitude tests used by Quantas against several pilot training criteria. Stead reported uncorrected correlations for three tests developed by the UK Royal Air Force (control of velocity, instrument comprehension, and sensorimotor apparatus) showing moderate validities ranging from .226 to .320 for small sample sizes between 186 and 234. Unfortunately, Stead only reported those correlations that were statistically significant. Further, he failed to correct for range restriction making it difficult to interpret these results. These correlations are conservative estimates of the validity of the selection system.

Doat (1995) described a 4-stage process used by Air France for selecting pilots. The stages included measurement of 1) general knowledge (e.g., mathematics, physics, mechanics, English proficiency), 2) "height" (e.g., dial and table reading, geographic orientation, numeric reasoning), 3) psychomotor coordination (e.g., attention, concentration, multi-tasking, perceptual speed, vigilance), and 4) psychological evaluation (e.g., personality, interviews). Doat reported that the rate of attrition varied for Stage 1 depending on the number of applicants and was estimated for Stages 2-4 respectively as 60% (of those surviving Stage 1), 60% (of those surviving Stages 1-2), and 50% (of those surviving Stages 1-3). Thus, the overall pass rate for Stages2-4 was about 18% of those surviving Stage 1 (.6 x .6 x .5 = .18). Doat (1995) noted that the proportion of applicants qualifying on the Air France tests has increased from 1990 to 1995 leading to concerns about possible test compromise. To remedy this trend, Doat called for revamping the test battery. Doat proposed developing alternate forms of the tests and giving the same pre-test practice to all applicants. Development of alternate forms of the cognitive tests (general knowledge and "height") requires writing many items for each form and ensuring equivalence of the forms in terms of content and difficulty, but is fairly straightforward. Development of alternate forms of the psychomotor tests is more difficult, especially if the goal is to replicate the general and specific factors measured by the tests. Air France already gives pilot applicants the same opportunity to practice. However, the amount of time spent by applicants on practice varies, probably as a function of their motivation.

Bartram and Baxter (1996) conducted a validation of the Cathay Pacific Airways pilot selection program. Based on a prior job analysis, Cathay Pacific selection procedures were designed to measure six "areas of competence." These included 1) technical skill and aptitude, 2) judgment and problem solving, 3) written and oral communications, 4) social relations, personality, and compatibility with Cathay, 5) leadership/subordinate style, and 6) motivation and ambition. Different selection procedures were used for cadets (i.e., *ab initio* pilot training applicants), second officers, and first officers. Likewise, separate validation studies were conducted for these three groups.

Cadets have a 3-stage selection process following a successful evaluation of their selection form. Stage 1 consists of aptitude testing (MICROPAT; Bartram, 1993), a test of English proficiency, an initial interview, and a short medical exam. Stage 2 includes personality assessment (16PF; Cattell, Eber, & Tatsuoka, 1970), a numerical reasoning test, a job knowledge test, a group-problem-solving-flight-planing exercise, general and technical interviews, and a full medical exam. Stage 3 consists of flying grading and a final selection board. Those accepted are offered a position in a training course.

For second and first officer applicants, selection consists of a 2-stage process following screening of their application. For those surviving the initial application screen, Stage 1 selection consists of an initial interview in their country of origin. Successful applicants are then invited to Hong Kong for Stage 2 which consists of separate general and technical interviews, a technical knowledge test, an assessment in the L1011 (Tri-star) full flight simulator, personality assessment (Cattell's 16 PF; Cattell, Eber, & Tatsuoka, 1970) and an extensive flight medical exam.

The performance criteria for *ab initio* students were training grades and final training outcome. For the second and first officer candidates, the criteria consisted of simulator and in-flight checks. Bartram and Baxter (1996) conducted analysis for each group of applicants and reported acceptable levels of validity for all three groups. It should be noted that the most valid selection measures varied by group. Results for the cadet sample are difficult to interpret due to the effects of range restriction caused by the multi-stage selection process and the small number of participants (n = 29) with training outcome data. For the cadets, the best predictors of training success were scores derived from the flight planning exercise and a conscientiousness score. The score based on aptitude and general ability showed poor validity. The finding for aptitude and general ability is surprising in light of their demonstrated utility for predicting pilot performance across several studies spanning many years (Ree & Carretta, 1996). For both second officers (n = 169) and first officers (n = 467), the best predictors of whether an applicant was selected for employment were the interviews and simulator ratings. These results of these studies may be misleading because they were not corrected for study artifacts such as range restriction and unreliability.

Two studies examined the validity of the DLR (German) pilot selection system for *ab initio* training at non-German applications, IBERIA Airlines (Stahlberg & Hörmann, 1993) and the Chinese Civil Aviation Flying College (Hörmann & Luo, 1999). Selection at IBERIA consists of five separate stages: 1) paper-and-pencil tests, 2) apparatus tests, 3) additional oral English exam, 4) medical exam, and 5) psychological interview. Eleven paper-and-pencil tests measure English proficiency, cognitive ability, and personality and three apparatus tests measure multiple task performance, psychomotor coordination, and choice reaction time. Historically, most IBERIA pilot applicants are screened out in the first two selection stages: paper-and-pencil tests (61% eliminated), apparatus tests (9%), English proficiency (4%), medical exam (2%), psychological interview (3%), and other (4%). As a result, about 16% of the applicants are accepted for training. The validation sample (Stahlberg & Hörmann, 1993) consisted of 98 student pilots who passed the selection screen. Training criteria included two written theoretical license exams, check flight scores, instructors' ratings, and a final pass/fail criterion. The paper-and-pencil aptitude tests showed their greatest validity against the written private pilot's license

total score. Overall, the apparatus tests measuring psychomotor ability and multiple task performance showed the greatest utility for prediction of performance. Similar results were reported for 125 Chinese student pilots attending a civil aviation flying college (Hörmann & Luo, 1999). Although the results from these two studies are informative, the small sample sizes and the lack of correction for range restriction and unreliability make the results potentially misleading.

Hörmann and Maschke (1996) examined the predictive validity of personality measures for airline pilot performance in the presence of other selection instruments. Participants were 274 licensed airline pilots to be employed by a European charter airline. Total flying hours varied from 150 to 19,100 hours, with a mean of 6,695 hours. Ninety-five percent of the sample had at least 1,000 flying hours. On average, participants had 8.6 years of airline experience. Selection was based on several factors including an interview, biographical data, a simulator check flight, a multidimensional personality test (Temperament Structure Scales [TSS]; Goeters, Timmermann, & Maschke, 1993), and the Cockpit Management Attitudes Questionnaire (CMAQ; Helmreich, 1984)[iv]. The performance criterion was a dichotomous variable ("standard" or "below standard") that was collected after about three years of employment with the hiring airline. Job success was rated as "standard" when no appreciable negative performance was observed and was rated as "below standard" in instances where a pilot was dismissed or when more than one recheck or irregularity report was recorded. Eighty-four percent of the pilots were subsequently rated as "standard." Hörmann and Maschke examined three correlational models to determine the predictive validity and incremental validity of the variables. Model 1 (6 variables) included five flying experience scores (number of flying hours, number of years airline experience, command experience [Y/N], jet experience [Y/N], and type experience [Y/N]) and age. Model 2 (7 variables) included the Model 1 scores and a simulator check flight grade. Model 3 (15 variables) was Model 2 with the eight TSS scales[v] added. An examination of the means on the TSS scales indicated that the average profile of the "standard" group was more favorable than that of the "below standard" group. Statistical comparisons indicated that the "standard" group was more emotionally stable, empathic, and energetic and less aggressive than the "below standard" group.

Results of the regression analyses indicated that all three models were valid and statistically significant ($p < .05$) predictors of the criteria. The respective multiple correlations were .39, .47, and .53. Some of the flying experience variables had unexpected negative relationships indicating that pilots with less experience were predicted to perform better than those with more experience. Unfortunately, a correlation matrix was not provided, so it is unknown whether or not the correlations between the individual variables and the criterion were in the expected positive direction. It is likely that the negative regression weights were a function of the intercorrelation of the independent variables. This intercorrelation of the independent variables is known as multicolinearity (Devore & Peck, 1993).

The simulator check flight and TSS scores were related in the expected direction to the criterion, but the eight personality scores failed to show incremental validity beyond flying experience, age, and the check flight ($F(8, 258) = 2.28$, ns). A comparison between Model 2 and Model 1 showed a significant incremental contribution for the check flight grade beyond prior flying experience and age ($F(1, 266) = 20.11$, $p < .01$).

While it is clear that non-US carriers are concerned about having valid selection systems, it is also clear that their research and statistical methodology often lacks the sophistication needed. Because of this their results cannot add to our understanding of the best practical and scientific methods for selecting pilots.

Psychological evaluations for existing pilots. The use of psychological evaluations for commercial pilots is more formalized among European than among United States air carriers (Goeters, 1995). In 1991, psychological requirements for commercial pilots were submitted to the Flight Crew Licensing Medical Group (FCL MED) of the European Civil Aviation Conference and were subsequently adopted. The requirements state:

> The applicant for the holder of a Class I (Commercial) or Class II (Private) medical certificate shall have no established psychological deficiencies, particularly in operational aptitudes or any relevant personality factor which is likely to interfere with the safe exercise of the privileges of the applicable license(s) (Goeters, 1995, p. 149).

Goeters (1995), citing the ECAC draft guidelines stated that a complete psychological evaluation includes an assessment of biographical data, aptitude and personality tests, and a psychological interview. Examples of biographical data include general life history, family, work history, health, and others. Aptitudes include a variety of cognitive and psychomotor abilities. Personality factors include decision-making, motivation, social capability, stress coping, and work orientation. Guidelines for the structure and content of the interview were not provided.

When the psychological evaluation was implemented, it was not a part of routine medical examination, but was initiated when the Aeromedical Board received information that led to concerns about the aptitude or personality of the pilot. Although this psychological evaluation was not included in the initial Class I (Commercial) evaluation, some (Goeters, 1995) have suggested that it be included. It should be noted, however, that support for this type of testing is not universal (Johnston, 1996; Murphy, 1995). A major concern of commercial pilots (Murphy, 1995) seems to be that under the current regulations, the decision as to whether a pilot is "operationally fit" is made by aviation psychologists or physicians, not other pilots. Johnston (1996) reviewed the arguments for and against the psychological testing of European pilots. Although Johnston noted that psychological testing might provide some economic and training benefits (e.g., reduction in training attrition), he concluded that such testing has many problems and risks. These include a possible shortage of qualified aviation psychologists to perform the assessments, a lack of an accepted test battery or standards of "acceptable performance," concerns with the psychometric properties (i.e., reliability, validity) of pilot assessment tests, cultural differences affecting test performance and interpretation of scores, and others. Upon reviewing the Joint Aviation Authorities' (JAA) guidance for psychological assessment, Johnston stated:

> The proposal that such tests and criteria are suitable for assessing the "psychological fitness" of experienced pilots is felt by many observers to be bizarre. It certainly appears to go well beyond the available evidence regarding the limitations and predictive capabilities of these tests. (p. 190)

Summary

The most common selection methods for commercial air carriers vary by type of carrier. However, interviews, background and reference checks, prior flying experience, and hands-on flying performance are common. Aptitude and personality measures are less common. In the few instances where validity studies are reported, it is difficult to interpret the results due to a variety of methodological problems (e.g., low statistical power, failure to correct for statistical artifacts) and a failure to report results in sufficient detail (i.e., means, standard deviations, and correlations of variables for training/job applicants and those accepted for training/employment). However, most of the reported studies provide conservative estimates of the value of the selection system.

Publication of future studies involving commercial pilot selection is strongly encouraged for two reasons. First, to share information on predictiveness of selection methods for commercial pilot selection and thus improve selection methods. Second, to provide sufficient detail such as means, variances, and correlation coefficients of all variables to allow secondary analyses (e.g., meta-analysis; Hunter & Schmidt, 1990). The ability to pool data across multiple studies, correct for study artifacts such as range restriction, sampling error, and unreliability allows researchers to obtain better estimates of the effectiveness (i.e., predictive validity) of selection methods.

The role of psychological assessment for licensing of commercial pilots is controversial. Proponents see it as a means of identifying psychological deficits of pilots and reducing potential risks to aviation safety. Opponents of psychological assessment for licensing express fears of abuse and concerns with the use of tests in circumstances for which they were not designed. Clearly, this is an area that will receive attention from aviation psychologists, pilots, and aviation industry representatives for some time.

THE FUTURE OF PILOT SELECTION METHODS

General Cognitive Ability

Because general cognitive ability has been demonstrated to be a versatile predictor in pilot selection, we discuss some emerging methods. Three noteworthy methods in the measurement of general cognitive ability that might be applied to the selection of pilots are chronometric measures, neural conductive velocity, and cognitive components.

Chronometric measures are typified by reaction time and choice reaction time. Jensen (1980, 1998) has shown that simple reaction time is correlated with measures of intelligence (i.e., *g*). This simple reaction-time task in which a finger is placed on a "home" button and moved to a "target" button when a light comes on, shows a low but positive correlation with measured intelligence. Choice reaction time, which requires pressing the one lighted button among many (for example eight), shows moderate correlation with measured intelligence.

The speed at which a neuron transmits an impulse is called neural conductive velocity. It requires no physically invasive procedure and is typically measured in the optic nerve. Electrodes are attached to the participant's head and a light is flashed which the participant sees. The verbal instructions are as simple as "look at the light" and no overt response is required by the participant. The head-mounted electrodes are connected to a computer with special software that measures the speed of the nerve impulse. Reed and Jensen (1992) have shown that neural conductive velocity in the optic nerve is correlated a moderate .37 with measured intelligence.

Cognitive components such as information processing speed and working memory capacity have been shown to be predominately measures of *g* (Kranzler & Jensen, 1991; Kyllonen & Christal, 1990; Miller & Vernon, 1992). The measurement of cognitive components is frequently done with computers. For example, participants may be shown a series of letters and sequentially told a set of rules governing the order of the letters. Next following the rules, participants must state the proper order of the letters. Arthur, Barrett, and Doverspike (1990) have shown validity for these types of tests in an occupational setting. However, Stauffer, Ree, and Carretta (1996) have demonstrated that cognitive components tests measure mostly *g* so that no major improvements in validity can be expected. Not withstanding, Seamster, Redding, and Kaempf (1997) have speculated that cognitive task analysis could lead to identification of specific cognitive components that would be predictive of pilot performance. This is inconsistent with past results and their speculations are highly unlikely to be fruitful as research results have shown that cognitive components measure mostly *g* (Stauffer et al., 1996). Further, Jones and Ree (1998) and Schmidt, Hunter, and Pearlman (1981) have demonstrated that job task differences do not change the effectiveness of general cognitive ability as a predictor. Empirical results to support the speculations of Seamster et al. would be helpful.

These three methods, chronometric measures, neural conductive velocity, and cognitive components may be fruitful only because they measure mostly *g*. They also offer the advantage of being content-free, thereby potentially reducing mean test score differences that were the consequences of differential educational choices. With careful psychometric development yielding increased reliability, these measures could find a place in pilot selection.

Flying Knowledge and Skills

Another potential trend in pilot selection could emerge in the measurement of flying knowledge and skills. As previously discussed, although simulation-based tests of flying skills (Gress & Willkomm, 1996; Okros et al., 1991) have shown validity for pilot selection, their use is relatively rare due to development and operating costs and the need for centralized administration. Advances in computer hardware and software have resulted in progressively faster, smaller, and less expensive computers, that in turn have made computer-based tests more common (Ree & Carretta, 1998). In recent applications, the US Air Force has developed several experimental "work sample" tests for possible use in selection of pilots and other aircrew specialties (i.e., air traffic controllers, navigators, and weapons directors). These tests resemble some aspect of job performance (e.g., instrument flight, cross-check) and require learning and applying complex rules.

Although these computerized tests exhibit face validity, their utility for pilot selection has yet to be determined.

Incrementing the Validity of g

Finally, the seminal work of Schmidt and Hunter (1998) that showed the job related validity of *g* and incremental validity of other predictors should serve as a guide to future research and practice. They collected hundreds of validation studies involving training and job performance and noted the predictors used. Predictors included *g*, selection interviews, work samples, personality, etc. Then they conducted numerous meta-analyses, one for each type of predictor when it was found to be used in conjunction with *g*. They computed the overall meta-analytic correlation of *g* with performance and then the meta-analytic multiple correlation of *g* and the other predictor with performance. Based on these meta-analytic findings, they concluded that the three combinations of predictors with the highest validity and utility for job performance were *g* plus a work sample test, *g* plus an integrity test (or a conscientiousness test), and *g* plus a structured interview. They also concluded that the latter two methods were appropriate for both entry-level selection and selection of experienced employees.

SUMMARY

Hunter and Burke (1995) offer a practical summary of pilot selection. They have created a how-to manual that embodies the current science and art of selection. Although we cannot agree with them on all details (e.g., their failure to correct studies for statistical artifact), their book contains many practical ideas and should be studied by all concerned with pilot selection.

The military has a long history of research in the selection of pilots and other aviation occupations. In general, they have used both paper-and-pencil tests and apparatus tests such as psychomotor. Cumulative results suggest that general cognitive ability (*g*) has been a mainstay of military testing and will likely remain so. Measures of pilot job knowledge and psychomotor ability have demonstrated incremental validity when used with measures of *g*.

American law requires job analyses for the development of job selection tests. The results of the analyses should be converted to good practice guided by cumulative knowledge. There is no single ideal pilot selection system, because not all pilots are hired the same way. Some are hired directly from the military with many flying hours, some from other airlines, and some directly from training. Although different, all the selection systems should be expected to have three common measurement elements: cognitive ability, conscientiousness (or possibly "integrity"), and job knowledge (Schmidt & Hunter, 1998).

There is a dearth of studies reported by American commercial airlines. Most likely, this is a consequence of two factors: legal liability and competitive edge. Results

from two recent surveys by the FAA suggest that US commercial airlines rely heavily on recruiting applicants with prior pilot experience. Prior experience can be assessed in many ways including background checks, interviews, examination of logbooks, flight simulators, and check flights. Aptitude and personality testing have received relatively little emphasis. In the instances where airlines employ *ab initio* selection (e.g., Bartram & Baxter, 1996), test batteries similar to those commonly found in military pilot selection are used.

The role of psychological evaluation in the licensing of airline pilots has been raised and debated in Europe. Proponents of psychological assessment for licensing see it as a means of identifying psychological deficits of pilots and reducing potential risks to aviation safety. Opponents express fears of abuse and concerns with the use of tests in circumstances for which they were not designed. Clearly, this is an area that will receive attention from aviation industry representatives, aviation psychologists, and pilots for some time.

There has been little use of personality assessment in the United States and the United Kingdom. Personality assessment is more prevalent in continental Europe and the cumulative research suggests that incremental validity could be achieved by using measures of personality, particularly conscientiousness (Barrick & Mount, 1991) or integrity (Schmidt & Hunter, 1998).

Most important in conducting pilot selection research is scientific rigor. Without scientific rigor, results may be worse than meaningless leading to counterproductive practice. Before setting out to develop a pilot selection system, it is imperative to have a firm foundation in the published literature of human abilities, reliability, validity, job performance measurement, and meta-analysis. Cumulative research results should guide practice.

BIBLIOGRAPHICAL NOTE

Thomas R. Carretta received his Ph.D. in psychology in 1983 from the University of Pittsburgh. Currently, he is a Research Psychologist in the Crew Systems Development Branch of the Human Effectiveness Directorate of the Air Force Research Laboratory in Dayton, Ohio. Prior to his current position, he spent over 10 years in the Manpower and Personnel Research Division of the Air Force Research Laboratory in San Antonio, Texas working on aircrew selection and classification issues. His professional interests include personnel measurement, selection, and individual and group differences.

Malcolm James Ree is associate professor at the Center for Leadership Studies at Our Lady of the Lake University, San Antonio, TX. Formerly, he was the Senior Scientist in the Space Warfighter Training Branch of the Human Effectiveness Directorate of the Air Force Research Laboratory. He received his Ph.D. in psychometrics in 1976 from the University of Pennsylvania. His professional interests include human abilities, individual differences, the intelligence-job performance nexus, and statistics.

ACKNOWLEDGEMENTS

The authors offer special thanks to Dr. David R. Hunter for his help in locating materials regarding commercial pilot selection practices. We also thank Dr. Monica Martinussen, Eugene Burke, and the book editors for their comments on a previous draft. Finally we thank Mark Bowler, Rui Bartolo Ribeiro, Hans-Jürgen Hörmann , Jose Puente, Paul Rioux, James Steindl, and Dr. Joseph Weeks for their help in this effort.

Please send correspondence for Dr. Thomas R. Carretta to AFRL/HECI, 2210 8th Street, Bldg. 146, Room 122, Wright-Patterson AFB, OH 45433-7511. Send e-mail to thomas.carretta@wpafb.af.mil. Send correspondence for Dr. Malcolm James Ree to 411 S. W. 24th Street, Center for Leadership Studies, Moye 410, San Antonio Texas 78207, e-mail mree@stic.net.

REFERENCES

Aiken, L. R., Jr. (1966). Another look at weighting test items. *Journal of Educational Measurement, 3,* 183-185.

Alwin, D. F., Jackson, D. J. (1981). Application of simultaneous factor analysis to issues of factor invariance. In D. J. Jackson & E. F. Borgatta (Eds.), *Factor analysis and measurement in sociological research* (pp. 249-279). Beverly Hills, CA: Sage.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (Joint Committee) (1985). *Standards for educational & psychological testing.* Washington, DC: American Psychological Association.

Arthur, W., Jr., Barrett, G. V., & Doverspike, D. (1990). The validation of an information-processing-based test battery for the prediction of handling accidents among petroleum product transport drivers. *Journal of Applied Psychology, 75,* 621-628.

Bailey, M., & Woodhead, R. (1996). Current status and future developments of RAF aircrew selection. *Selection and Training Advances in Aviation: AGARD Conference Proceedings 588* (pp. 8-1-8-9). Prague, Czech Republic: Advisory Group for Aerospace Research & Development.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance. *Personnel Psychology, 44,* 1-26.

Bartolo Ribeiro, R., (1992). Predicao da performance em psicologia aeronautica: Validacao de uma bateria de seleccao. *Analise Psicologia, Serie X, 3,* 353-365.

Bartolo Ribeiro, R., Martins, A., Vicoso, A., Carpinteiro, M. J., & Estrela, R. (1992). Validacaocapacidade preditiva dos testes utilizados na seleccao de pilotos. *Revista depsicologiamilitar, Numero Especial,* 271-280.

Bartram, D. (1993). Aptitude testing and selection in aviation. In R. Telfer (Ed.), *Aviation training and instruction* (pp. 34-51). Aldershot, England; Gower Ashgate.

Bartram, D., & Baxter, P. (1996). Validation of the Cathay Pacific Airways pilot selection program. *The International Journal of Aviation Psychology, 6,* 149-169.

Bennett, G. J., Seashore, H. G., & Wesman, A. G. (1982). *Differential Aptitude Tests (Forms V and W): Administrators Handbook.* New York: Psychological Corporation.

Bentler, P. M. (1989). *EQS structural equation program manual.* Los Angeles, CA: BMDP Statistical Software.

Boer, L. C. (1992). *TASKOMAT: Meaning of test scores.* Soesterberg, Netherlands: TNO Institute for Perception.

Boer, L. C., Harsveld, M., & Hermans, P. H. (1997). The selective-listening task as a test for pilots and air traffic controllers. *Military Psychology, 9,* 137-149.

Brand, C. (1987). The importance of general intelligence. In S. Modgil & C. Modgil (Eds.). *Arthur Jensen: Consensus and Controversy.* New York:: Falmer Press.

Burke, E. (1993). Pilot selection in NATO: An overview. In R. S. Jensen & D. Neumeister (Eds.) *Proceedings of the Seventh International Symposium on Aviation Psychology* (pp. 373-378). Columbus, OH: Ohio State University.

Burke, E. F. (1995). Male-female differences on aviation selection tests: Their implications for research and practice. In N. Johnston, R. Fuller, & N. McDonald (Eds.), *Aviation psychology: Training and selection* (pp. 188-193). Aldershot, England: Avebury Aviation.

Burke, E., Hobson, C., & Linsky, C. (1997). Large sample validations of three general predictors of pilot training success. *The International Journal of Aviation Psychology, 7,* 225-234.

Callister, J. D., King, R. E., Retzlaff, P. D., Marsh, R. W. (1999). Revised NEO Personality Inventory profiles for male and female U. S. Air Force pilots. *Military Medicine, 164,* 885-890.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmidt, W. C. Bowman, & Associates (Eds.). *Personnel selection in organizations* (pp. 35-70). San Francisco, CA: Josey-Bass.

Campbell, J. P., McHenry, J. J., & Wise, L. L. (1990). Modeling job performance in a population of jobs. Special Issue: Project A: The US Army Selection and Classification Project. *Personnel Psychology, 43,* 313-333.

Carretta, T. R. (1989). USAF pilot selection and classification systems. *Aviation, Space, and Environmental Medicine, 60,* 46-49.

Carretta, T. R. (1990). Cross-validation of experimental USAF pilot training performance models. *Military Psychology, 2,* 257-264.

Carretta, T. R. (1992a). Recent developments in U. S. Air Force pilot candidate selection and classification. *Aviation, Space, and Environmental Medicine, 63,* 1112-1114.

Carretta, T. R. (1992b). Understanding the relations between selection factors and pilot training performance: Does the criterion make a difference? *The International Journal of Aviation Psychology, 2,* 95-105.

Carretta, T. R. (1997a). Group differences on US Air Force pilot selection tests. *International Journal of Selection and Assessment, 5,* 115-127.

Carretta, T. R. (1997b). Male-female performance on U. S. Air Force pilot selection tests. *Aviation, Space, and Environmental Medicine, 68,* 818-823.

Carretta, T. R. (2000). US Air Force pilot selection and training methods. *Aviation, Space, and Environmental Medicine, 71,* xxx-yyy.

Carretta, T. R., Hansen, I., & Woodhead, R. W. A. (2000). *Euro-NATO Aircrew Human Factors Working Group (AHFWG) Activities overview: 1982-1999,* Euro-NATO-AHFWG-TR-05. Euro-NATO Aircrew Human Factors Working Group.

Carretta, T. R., Perry, D. C., Jr., & Ree, M. J. (1996). Prediction of situational awareness in F-15 pilots. *The International Journal of Aviation Psychology, 6,* 21-41.

Carretta, T. R., & Ree, M. J. (1993). Basic Attributes Test (BAT): Psychometric equating of a computer-based test. *The International Journal of Aviation Psychology, 3,* 189-201.

Carretta, T. R., & Ree, M. J. (1994). Pilot candidate selection method (PCSM): Sources of validity. *The International Journal of Aviation Psychology, 4,* 103-117.

Carretta, T. R., & Ree, M. J. (1995a). Air Force Officer Qualifying Test Validity for predicting pilot training performance. *Journal of Business and Psychology, 9,* 379-388.

Carretta, T. R., & Ree, M. J. (1995b). Near identity of cognitive structure in sex and ethnic groups. *Personality and Individual Differences, 19,* 149-155.

Carretta, T. R., & Ree, M. J. (1996a). Factor structure of the Air Force Officer Qualifying Test: Analysis and comparison. *Military Psychology, 8,* 29-42.

Carretta, T. R., & Ree, M. J. (1996b). U. S. Air Force pilot selection tests: What is measured and what is predictive? *Aviation, Space, and Environmental Medicine, 67,* 275-283.

Carretta, T. R., & Ree, M. J. (1997). A preliminary evaluation of causal models of male and female acquisition of pilot skills. *The International Journal of Aviation Psychology, 7,* 353-364.

Carretta, T. R., & Ree, M. J. (in press). Pitfalls of ability research. *International Journal of Selection and Assessment.*

Carretta, T. R., Rodgers, M. N., & Hansen, I. (1993). *The identification of ability requirements and selection instruments for fighter pilot training*, AL-TP-1993-0016. Brooks Air Force Base, TX: Armstrong Laboratory, Human Resources Directorate, Manpower and Personnel Research Division.

Cascio, W. F. (1982). *Costing human resources: The financial impact of behavior in organizations*. New York: Van National Reinhold.

Cascio, W. F. (1991). *Applied psychology in personnel management* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire (16PF)*. Champaign, IL: Institute for personality and Ability Testing.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249-253.

Cohen, J. (1987). *Statistical power analysis for the behavioral sciences* (rev. ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual.* Odessa, FL: Psychological Assessment Resources, Inc.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Damos, D. L. (1996). Pilot selection batteries: Shortcomings and perspectives. *The International Journal of Aviation Psychology, 6,* 199-209.

Dansereau, F., Alutto, J. A., & Yammarino, F. J. (1984). *Theory testing in organizational behavior: The variant approach.* Englewood Cliffs, NJ: Prentice-Hall.

Delaney, H. D. (1992). Dichotic listening and psychomotor task performance as predictors of naval primary flight-training criteria. *The International Journal of Aviation Psychology, 2,* 107-120.

Devore, J., & Peck, R. (1993). *Statistics: The exploration and analysis of data* (2nd ed.). Belmont, CA: Duxbury.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41,* 417-440.

Doat, B. (1995). Need of new development in Air France selection. In N. Johnston, R. Fuller, & N. McDonald (Eds.), *Aviation psychology: Training and selection. Proceedings of the*

21st Conference of the European Association for Aviation Psychology (EAAP), Vol. 2 (pp. 182-187). Aldershot, England: Avebury Aviation.

Dockeray, F. C., & Isaacs, S. (1921). Psychological research in aviation in Italy, France, England, and the American Expeditionary Forces. *Journal of Comparative Psychology, 1,* 115-148.

Dolgin, D. L., & Gibb, G. D. (1989). Personality assessment in aviator selection. In R. S. Jensen (Ed.), *Aviation psychology* (pp. 288-320). London: Gower Publishing.

Duke, A. P., & Ree, M. J. (1996). Better candidates fly fewer training hours: Another time testing pays off. *International Journal of Selection and Assessment, 4,* 115-121.

Dvorak, B. J. (1947). The new U. S. E. S. General Aptitude Test Battery. *Journal of Applied Psychology, 31,* 372-376.

Earles, J. A., & Ree, M. J. (1992). The predictive validity of ASVAB for training grades. *Educational and Psychological Measurement, 52,* 721-725.

Equal Employment Opportunity Commission (1978). *Uniform guidelines on employee selection procedures.* Title 29- Labor, Part 1607. *National Archives and Records Administration code of federal regulations.* Washington, DC: U. S. Government Printing Office.

Evdokimov, V. L. (1988). Socio-psychological selection in the system of professional training of pilots. *Psikocheskii Zhurnal, 9,* 71-74.

Fiske, D. W. (1947). Validation of naval aviation cadet selection tests against training criteria. *Journal of Applied Psychology, 31,* 601-614.

Flanagan, J. C. (1942). The selection and classification program for aviation cadets (aircrew-bombardiers, pilots, and navigators). *Journal of Consulting Psychology, 5,* 229-238.

Fuller, W. A. (1987). *Measurement error models.* New York: Wiley.

Gael, S. (1988). *The job analysis handbook for business, industry, and government.* Vols. 1 and 2. New York: Wiley.

Geldard, F. A., & Harris, C. W. (1946). Selection and classification of aircrew by the Japanese. *American Psychologist, 1,* 205-217.

Gibb, G. D., & Dolgin, D. L. (1989). Predicting military flight training success by a compensatory tracking task. *Military Psychology, 1,* 235-240.

Goeters, K. M. (1995). Psychological evaluation of pilots: The present regulations and arguments for their application. *Aviation psychology: Training and selection. Proceedings of the*

21st Conference of the European Association for Aviation Psychology (EAAP), Vol. 2 (pp. 149-156). Aldershot, England: Avebury Aviation.

Goeters, K. M., Timmermann, B., & Maschke, P. (1993). The construction of personality questionnaires for selection of aviation personnel. *The International Journal of Aviation Psychology, 3,* 123-141.

Goldberg, L. R. (1992). The development of markers of the Big Five factor structure. *Psychological Assessment, 4,* 26-42.

Goldberg, S. (1991). *When wish replaces thought.* Buffalo, New York: Prometheus.

Gopher, D., & Kahneman, D. (1971). Individual differences in attention and the prediction of flight criteria. *Perceptual and Motor Skills, 33,* 1334-1342.

Gress, W., & Willkomm, B. (1996). Simulator-based test systems as a measure to improve the prognostic value of aircrew selection. *Selection and Training Advances in Aviation: AGARD Conference Proceedings 588* (pp. 15-1-15-4). Prague, Czech Republic: Advisory Group for Aerospace Research & Development.

Guion, R. M. (1976). Recruiting, selection, and job placement. In M. D. Dunnett (Ed.) *Handbook of industrial and organizational psychology* (pp. 777-828). Chicago: Rand McNally.

Gulliksen, H., & Wilks, S. S. (1950). Regression tests for several samples. *Psychometrika, 15,* 91-114.

Hansen, I. (1999). *Use of interviews by Euro-NATO air forces in pilot selection* (Tech. Rep. No. AHFWG-3). Euro-NATO Aircrew Human Factors Working Group.

Hansen, J. S., & Oster, C. V., Jr. (Eds.). (1997). *Taking flight: Education and training in aviation careers.* Washington, DC: National Academy Press.

Harsveld, M. (1991). The Defense Mechanism Test and success in flying training. In E. Palmer (Ed.), *Human resource management in aviation.* Aldershot, UK: Avebury Technical.

Helmreich, R. L. (1984). Cockpit management attitudes. *Human Factors, 26,* 583-589.

Hilton, T. F., & Dolgin, D. L. (1991). Pilot selection in the military of the free world. In G. Gal & A. D. Mangelsdorff (Eds.), *Handbook of military psychology* (pp. 81-101). New York: Wiley.

Hogan, J. C. (1991). Physical abilities. In M. N. Dunnette & L. M. Hough (Eds.). *Handbook of industrial and organizational psychology* (2nd ed.) Vol. 2 (pp. 753-871). Palo Alto, CA: Consulting Psychologists Press.

Hörmann, H. J., & Luo, X. L. (1999). Development and validation of selection methods for Chinese student pilots. In R. S. Jensen, B. Cox, J. D. Callister, & R. Lavis (Eds.), *Proceedings of the Tenth International Symposium on Aviation Psychology* (pp. 571-576). Columbus, OH: Ohio State University.

Hörmann, H. J., & Maschke, P. (1991). Exogenous and endogenous determinants of cockpit management attitudes. In R. S. Jensen (Ed.), *Proceedings of the Sixth International Symposium on Aviation Psychology* (pp. 384-3390). Columbus, OH: Ohio State University.

Hörmann, H. J., & Maschke, P. (1996). On the relations between personality and job performance of airline pilots. *The International Journal of Aviation Psychology, 6,* 171-178.

Houck, M. R., Whitaker, L. A., & Kendall, R. R. (1991). *Behavioral taxonomy for air combat F-15 defensive counter-air mission*_(UND-TR-91-147). Dayton, OH: University of Dayton Research Institute.

Hunter, D. R. (1989). Aviator selection. In M. S. Wiskoff & G. M. Rampton (Eds.), *Military personnel measurement: Testing, assignment, evaluation* (pp. 129-167). New York: Praeger.

Hunter, D. R., & Burke, E. F. (1994). Predicting aircraft pilot training success: A meta-analysis of published research. *The International Journal of Aviation Psychology, 4,* 297-313.

Hunter, D. R., & Burke, E. F. (1995). *Handbook of pilot selection.* Brookfield, VT: Avebury Aviation.

Hunter, J. E. (1980). *The dimensionality of the General Aptitude Test Battery (GATB) and the dominance of general factors over specific factors in the prediction of job performance.* Washington DC: US Employment Service, UD Department of Labor.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29,* 340-362.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors across studies. *Psychological Bulletin, 96,* 72-98.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis.* Newbury Park, CA: Sage.

Imhoff, D. F., & Levine, J. M. (1981). *Perceptual-motor and cognitive performance task battery for pilot selection* (Tech. Rep. No. AFHRL-TR-80-27). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Jensen, A. R. (1980). *Bias in mental testing.* New York: The Free Press.

Jensen, A. R. (1993). Commentary: Vehicles of *g*. *Psychological Science, 3,* 275-278.

Jensen, A. R. (1994). What is a good *g*? *Intelligence, 18,* 231-258.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.

Jessup, G., & Jessup, H. (1966). Validity of the Eyesenck Personality Inventory for pilot selection. *Journal of Occupational Psychology, 45,* 111-123.

Johnson, J. T., & Ree, M. J. (1994). RANGEJ: A Pascal program to compute the multivariate correction for range restriction. *Educational and Psychological Measurement, 54,* 693-695.

Johnston, N. (1996). Psychological testing and pilot licensing. *The International Journal of Aviation Psychology, 6,* 179-197.

Jones, G. E., & Ree, M. J. (1998). Aptitude test validity: No moderating effects due to job ability requirements. *Educational and Psychological Measurement 58,* 282-292.

Jöreskog, K., & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago, IL: Scientific Software International.

Kennedy, E. (1988). Estimation of the squared cross-validity coefficients in the context of best subtest regression. *Applied Psychological Measurement, 12,* 231-237.

Kim, J. O., & Mueller, C. W. (1988). *Factor analysis: Statistical methods and practical issues.* Beverly Hills, CA: Sage Publications.

Kranzler, J. H., & Jensen, A. R. (1991). The nature of psychometric *g*: Unitary process or a number of independent processes? *Intelligence, 15,* 397-422.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working memory capacity?! *Intelligence, 14,* 389-433.

Lance, C. L., Teachout, M. S., & Donnelly, T. M. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. *Journal of Applied Psychology, 77,* 437-452.

Lautenschlager, G. J., & Mendoza, J. (1986). A step-down hierarchical multiple regression analysis for estimating hypotheses about test bias in prediction. *Applied Psychological Measurement, 10,* 133-159.

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh, Section A, 62, Part 1,* 28-30.

Li, L. (1993). On military pilot selection. *Psychological Science (China), 16,* 299-304.

Linn, R. L., Harnish, D. L., & Dunbar, S. (1981). Corrections for range restriction: An empirical investigation of conditions resulting in conservative corrections. *Journal of Applied Psychology, 66,* 655-663.

Long, G. E., & Varney, N. C. (1975). *Automated pilot aptitude measurement system* (AFHRL-TR-75-58). Lackland AFB, TX: Air Force Human Resources Laboratory, Personnel Research Division.

Manzey, D., Hörmann, H. J., Osnabrügge, G., & Goeters, K. M. (1990). *International application of the DLR test-system: First year of cooperation with IBERIA in pilot selection* (DLR-FB-90-05). Hamburg. Germany: DLR Institut für Flugmrdizin, Abteilung Luft-und Raumfahrtpsychologie.

Martinussen, M. (1996). Psychological measures as predictors of pilot performance: A meta-analysis. *The International Journal of Aviation Psychology, 6,* 1-20.

Martinussen, M. (1997). Pilot selection and range restriction: A red herring or a real problem? *Proceedings of the Ninth International Symposium on Aviation Psychology* (pp. 1314-1318). Columbus, OH: Ohio State University.

Martinussen, M., & Torjussen, T. (1993). Does DMT (Defense Mechanism Test) predict pilot performance only in Scandinavia? *Proceedings of the Seventh International Symposium on Aviation Psychology* (pp. 398-403). Columbus, OH: Ohio State University.

Martinussen, M., & Torjussen, T. (1998). Pilot selection in the Norwegian Air Force: A validation and meta-analysis of the test battery. *The International Journal of Aviation Psychology, 8,* 33-45.

McArdle, J. J. (1996). Current directions in structural factor analysis. *Current Directions in Psychological Science, 5,* 11-18.

McClelland, D. C. (1993). Intelligence is not the best predictor of job performance. *Current Directions in Psychological Science, 2,* 5-6.

McCormick, E. J. (1976). Job and task analysis. In M. D. Dunnett (Ed.) *Handbook of industrial and organizational psychology* (pp. 651-696). Chicago: Rand McNally.

McCormick, E. J. (1979). *Job analysis: Methods and applications.* New York: AMACOM.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43,* 335-354.

McNemar, Q. (1969). *Psychological statistics* (4th ed.). New York: Wiley.

Melton, A. W. (Ed.). (1947). *Army Air Forces aviation psychology research reports: Apparatus tests* (Report No. 4). Washington, DC: GPO.

Messick, S. (1989). Validity. In R. L. Linn (Ed.). *Educational Measurement* (3rd ed., pp. 13- 103). New York: Macmillan.

Miller, L. T., & Vernon, P. A. (1992). The general factor in short-term memory, intelligence, and reaction time. *Intelligence, 16,* 5-29.

Morrison, T., & Morrison, M. (1995). A meta-analytic assessment of the predictive validity of the quantitative and verbal composites of the Graduate Record Examination with graduate grade point average representing the criterion of graduate success. *Educational and Psychological Measurement, 55,* 309-316.

Mosier, C. I. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement, 11,* 5-11.

Murphy, K. R. (1983). Fooling yourself with cross-validation: Single-sample designs. *Personnel Psychology, 36,* 111-118.

Murphy, T. (1995). JAA psychological testing of pilots: Objections and alarms. *Aviation psychology: Training and selection. Proceedings of the 21st Conference of the European Association for Aviation Psychology (EAAP), volume 2* (pp. 157-163). Aldershot, England: Avebury Aviation.

Novis Soto, M. L. (1998). Los cuestionarios de personalidad en la selection de los pilotos de línea aérea. *Revista de Psicologia del Trabajo y de las Organizaciones, 14,* 113-128.

Okros, A. C., Spinner, B., & James, J. A. (1991). *The Canadian Automated Pilot Selection System* (Research Report 91-1). Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.

Olea, M. M., & Ree, M. J. (1994). Predicting pilot and navigator criteria: Not much more than *g*. *Journal of Applied Psychology, 79,* 845-851.

Orgaz, B., & Loro, P. (1993). La psicologia aeronautico-militar en Espana: breve apunte historico. *Revista de Historia de la Psicologia, 14,* 271-283.

Parry, J. B. (1947). The selection and classification of RAF aircrew. *Occupational Psychology, 21,* 158-167.

Patrick, J. (in press). Training. In B. H. Kantowitz (Series Ed.) & P. S. Tsang & M. A. Vidulich (Vol. Eds.). *Human factors in transportation: Vol. xx. Principles and practices of aviation psychology* (pp. xxx-yyy). Mahwah, NJ: Erlbaum.

Prieto, G., Carro, J., Palenzuela, D. L., Pulido, R. F., Orgaz, B., Delgado, A. R., & Loro, P. (1996). Diferencias individuales y práctica profesional en el ámbito militar: Seleccion de pilotos áereos. In: M. DeJuan-Espinosa, R. Colom, & M. A. Quiroga, (Eds.), *La práctica de la psicologia diferencial en industria y organizaciones* (pp. 171-194). Madrid: Piramide.

Pulakos, E. (1997). Ratings of job performance. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology* (pp. 291-318). Palo Alto, CA: Davies-Black Publishing.

Raven, J. C. (1966). *Advanced Progressive Matrices.* New York: Psychological Corporation.

Ree, M. J. (1995). Nine rules for doing ability research wrong. *The Industrial-Organizational Psychologist, 32,* 64-68.

Ree, M. J., & Carretta, T. R. (1994a). Factor analysis of ASVAB: Confirming a Vernon-like structure. *Educational and Psychological Measurement, 54,* 457-461.

Ree, M. J., & Carretta, T. R. (1994b). The correlation of general cognitive ability and psychomotor tracking tests. *International Journal of Selection and Assessment, 2,* 209-216.

Ree, M. J., & Carretta, T. R. (1996). Central role of g in military pilot selection. *The International Journal of Aviation Psychology, 6,* 111-123.

Ree, M. J., & Carretta, T. R. (1997). What makes an aptitude test valid? In R. F. Dillon (Ed.). *Handbook on testing* (pp. 65-81). Westport, CT: Greenwood Press.

Ree, M. J., & Carretta, T. R. (1998). Computerized testing in the United States Air Force. *International Journal of Selection and Assessment, 6,* 82-89.

Ree, M. J., Carretta, T. R., & Earles, J. A. (1998). In top-down decisions, weighting variables does not matter: A consequence of Wilks' theorem. *Organizational Research Methods, 1,* 407-420.

Ree, M. J., Carretta, T. R., & Earles, J. A. (1999). In validation sometimes two sexes are one too many. *Human Performance, 12,* 79-88.

Ree, M. J., Carretta, T. R., Earles, J. A., & Albert, W. (1994). Sign changes when correcting for range restriction: A note on Pearson's and Lawley's selection formulas. *Journal of Applied Psychology, 79,* 298-301.

Ree, M. J., Carretta, T. R., & Teachout, M. S. (1995). Role of ability and prior job knowledge in complex training performance. *Journal of Applied Psychology, 80,* 721-730.

Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than *g*. *Personnel Psychology, 44,* 327-332.

Ree, M. J., & Earles, J. A. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science, 1,* 86-89.

Ree, M. J., & Earles, J. A. (1993). g is to psychology what carbon is to chemistry: A reply to Sternberg and Wagner, McClelland, and Calfee. *Current Directions in Psychological Science, 2,* 11-12.

Reed, T. E., & Jensen, A. R. (1992). Conduction velocity in a brain nerve pathway of normal adults correlates with intelligence level, *Intelligence, 16,* 259-272.

Retzlaff, P. D., & Gilbertini, M. (1987). Air force pilot personality: Hard data on the "right stuff." *Multivariate Behavioral Research, 22,* 383-399.

Reynolds, C. E. (1982). Methods for detecting construct and predictive bias. In R. A. Bork (Ed.) *Handbook of methods for detecting test bias* (pp. 199-277). Baltimore, MD: Johns Hopkins University Press.

Roberts. H. F., & Skinner. J. (1996). Gender and racial equity of the Air Force Officer Qualifying Test in officer training school selection decisions. *Military Psychology, 8,* 95-113.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology, 73,* 482-486.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22,* 53-61.

Schmidt, F. L., & Hunter, J. E. (1978). Moderator research and the law of small numbers. *Personnel Psychology, 31,* 215-232.

Schmidt, F. L., & Hunter, J. E. (1993). Tacit knowledge, practical intelligence, general mental ability, and job knowledge. *Current Directions in Psychological Science, 2,* 8-9.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274.

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1981). Task differences and validity of aptitude tests in selection: A red herring. *Journal of Applied Psychology, 66,* 166-185.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61,* 473-485.

Seamster, T. L., Redding, R. E., & Kaempf, G. L. (1997). *Applied cognitive task*

analysis in aviation. Brookfield, VT: Ashgate Publishing.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309-316.

Siem, F. M., Carretta, T. R., & Mercatante, T. A. (1988). *Personality, attitudes, and pilot training performance: Preliminary analysis* (AFHRL-TP-87-62). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Siem, F. M., & Murray, M. W. (1994). Personality factors affecting pilot combat performance: A preliminary investigation. *Aviation, Space, and Environmental Medicine, 65,* A45-A48.

Signori, E. I. (1949). The Arnprior experiment: A study of World War II pilot selection procedures in the RCAF and RAF. *Canadian Journal of Psychology, 3,* 136-150.

Skinner, J., & Ree, M. J. (1987). *Air Force Officer Qualifying Test (AFOQT): Item and factor analysis of Form O* (Tech. Rep. No. AFHRL-TR-86-68). Brooks Air Force Base e, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Smallwood, T., & Fraser, M. (1995). *The airline training pilot.* Brookfield, VT: Ashgate Publishing.

Society for Industrial-Organizational Psychology (1987). *Principles for the validation and use of personnel selection procedures* (3rd ed.). College Park, MD: Author.

Spearman, C. (1904). "General Intelligence," objectively determined and measured. *American Journal of Psychology, 15,* 201-293.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* New York: Macmillan.

Spinner, B. (1991). Predicting success in primary flying school from the Canadian Automated Pilot Selection System: Development and cross-validation. *The International Journal of Aviation Psychology, 1,* 163-180.

Stahlberg, G., & Hörmann, H. J. (1993). *International application of the DLR test-system: Validation of the pilot selection for IBERIA* (DLR-FB-93-42). Hamburg. Germany: DLR Institut für Flugmrdizin, Abteilung Luft-und Raumfahrtpsychologie.

Stauffer, J. M., Ree, M. J., & Carretta, T. R. (1996). Cognitive-components tests are not much more than g: An extension of Kyllonen's analyses. *The Journal of General Psychology, 123,* 193-205.

Stead, G. (1991). A validation study of the Quantas pilot selection process. In E. Farmer (Ed.), *Human resource management in aviation* (pp. 3-18). Aldershot, England: Avebury Technical.

Stead, G. (1995). Quantas pilot selection procedures: Past to present. In N. Johnston, R. Fuller, & N. McDonald (Eds.), *Aviation psychology: Training and selection. Proceedings of the 21st Conference of the European Association for Aviation Psychology (EAAP), Vol. 2* (pp. 176-181). Aldershot, England: Avebury Aviation.

Stein, C. (1960). Multiple regression. In I. Olkin et al. (Eds.), *Contributions to probability and statistics*. Stanford, CA: Stanford University Press.

Sternberg, R. J., & Wagner, R. K. (1993). The *g*-ocentric view of intelligence and job performance is wrong. *Current Directions in Psychological Science, 2,* 1-5.

Suarez, J., Barborek, S., Nikore, V., & Hunter, D. R. (1994). *Current trends in pilot hiring and selection* (Memorandum No. AAM-240-94-1). Washington, DC: Federal Aviation Administration.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991) Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44,* 703-742.

Thorndike, R. L. (1949). *Personnel selection*. New York: Wiley.

Thurstone, L. L. (1938). *Primary mental abilities*. Psychometric Monographs No 1.

Torjussen, T., & Vaemess, R. (1991). The use of the Defense Mechanism Test (DMT) in Norway for selection and stress research. In M. Olss, G. Dodaert, & H. Ursin (Eds.). *Quantification of human defense mechanisms*. Berlin: Springer-Berlag.

Tupes, E. C., & Christal, R. E. (1961). *Recurrent personality factors based on trait rankings* (Tech. Rep. No. ASD-TR-61-97). Lackland Air Force Base, TX: Personnel Laboratory, Aeronautical Systems Division.

Vernon, P. E. (1969). *Intelligence and cultural environment*. London: Methuen.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (1996). *Modeling job performance: Is there a general factor?* Manuscript submitted for publication.

Viteles, M. S. (1945). The aircraft pilot: 5 years of research. A summary of outcomes. *Psychological Bulletin, 42,* 489-526.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83,* 213-217.

Wainer, H. (1978). On the sensitivity of regression and regressors. *Psychological Bulletin, 85,* 267-273.

Walter, D. C. (1998). Air warriors: *The inside story of the making of a navy pilot.* New York: Simon & Schuster.

Walters, L. C., Miller, M., & Ree, M. J. (1993). Structured interviews for pilot selection: No incremental validity. *The International Journal of Aviation Psychology, 3,* 25-38.

Weeks, J. L., & Zelenski, W. E. (1998). *Entry to USAF undergraduate flying training* (AFRL-HE-AZ-TR-1998-0077). Brooks AFB, TX: Training Effectiveness Branch, Warfighter Training Research Division.

Weeks, J. L., Zelenski, W. E., & Carretta, T. R. (1996). Advances in USAF pilot selection. *Selection and Training Advances in Aviation* (AGARD-CP-588). Prague, Czech Republic, 1-1-1-11.

Wheeler, J. L., & Ree, M. J. (1997). The role of general and specific psychomotor tracking ability in validity. *International Journal of Selection and Assessment, 5,* 128-136.

Wherry, R. J. (1975). Underprediction from overfitting: 45 years of shrinkage. *Personnel Psychology, 29,* 1-18.

Whetzel, D. L., & McDaniel, M. A. (1997). Employment interviews. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology* (pp. 185-205). Palo Alto, CA: Davies-Black Publishing.

Whetzel, D. L., & Oppler, S. H. (1997). Validation of selection instruments. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology* (pp. 355-384). Palo Alto, CA: Davies-Black Publishing.

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika, 3,* 23-40.

Yerkes, R. M. (1919). Report of the psychology committee of the National Research Council. *Psychological Review, 26,* 83-149.

[i] At the time of this study, RAF Basic Flying Training (BFT) consisted of Elementary Flying Training (up to 65 hours) in the Bulldog aircraft followed by BFT (up to 120 hours) in the Tucano aircraft.

[ii] Discriminant analysis is useful for situations where it is desirable to build a predictive model of group membership based on observed characteristics of each case. The procedure generates a discriminant function based on linear combinations of the predictor variables that provides the best discrimination between the groups. If there are more than two groups, a set of discriminant functions is required. The functions are generated from a sample of cases for which group membership is known. The functions can then be applied to new cases with data for the predictor variables, but unknown group membership.

[iii] Adds to less than 33% due to rounding.

[iv] The CMAQ analyses are not reported here. They were discussed in a separate study (Hörmann & Maschke, 1991) that examined the relations between personality variables measured by the TSS and cockpit management attitudes measured by the CMAQ.

[v] The eight TSS scales are Extraversion, Dominance, Emotional Instability, Aggressiveness, Empathy, Achievement Motivation, Rigidity, and Vitality.